THE *YERSINIA PESTIS* AUTOTRANSPORTER YAPG CONTAINS A FAST FOLDING β-HELIX DOMAIN

Monica L. Frazier

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry and Biophysics.

Chapel Hill
2012

Approved by:

Matthew Redinbo, Ph.D.

Richard Wolfenden, Ph.D.

Brian Kuhlman, Ph.D.

Nikolay Dokholyan, Ph.D.

Kevin SIep, Ph.D.

Virginia Miller, Ph.D.

ABSTRACT

MONICA L. FRAZIER: The *Yersinia pestis* Autotransporter YapG Contains a Fast Folding β-helix Domain
(Under the direction of Matthew Redinbo)

Autotransporter proteins are the most widely secreted protein family in gram-negative bacteria; their passenger domains are predicted to be β-helical in 97% of cases. The β-helical fold has been hitherto understudied with respect to protein folding, which typically is centered on small α-helical, low contact order proteins. In contrast, the β-helical portions of passenger domains are typically large, and made up of unique structural repeats with high contact order. Here, we have studied the *in vitro* folding of the passenger domain of YapG, an autotransporter from *Yersinia pestis*, via thermodynamic and kinetic approaches. We have identified YapG as the fastest refolding passenger domain to date. Steady-state fluorescence and circular dichroism indicate a one-step folding process; however, stopped-flow fluorescence indicates a one-step unfolding and a two-step refolding. Neither proline isomerization nor aggregation is associated with YapG refolding, suggesting this fast folding β-helix may experience a general collapse followed by a slower fine tuning folding step. In addition, gel filtration studies of the refolded state indicate that YapG may refold into two different folded species. Taken together, these results provide the first biophysical analysis of an autotransporter passenger domain from *Y. pestis* and provide new insight into the folding process in β-helical folds.

Dedication

To my daughter, Katie Mabel Frazier, who I hope will chase her dreams no matter how big.

ACKNOWLEDGEMENTS

There are many people who should be acknowledged as I finish my dissertation. First, I must thank Dr. Paulo Almeida and Dr. Antje Almeida, who stimulated my love of research. My work with you in undergrad helped me learn how to use a curious mind in focused research. Your guidance led me to where I am today. Your work and your lives are an inspiration to me; you will always be a part of my family. In a surprise twist, I also got to work with Antje during my graduate studies. I must thank you both for your hospitality when I came back to Wilmington to do experiments. Antje, your interest in my project and intellectual and experimental support were instrumental in the finalization of my dissertation and for that I cannot thank you enough. In addition, I must acknowledge the entire Chemistry Department at the University of North Carolina at Wilmington, whose faculty does a fantastic job preparing their undergraduates for either the job market or graduate school. Entering into biophysics at UNC I was worried about my preparation in comparison to my peers, but as I found out, there was no need to worry.

I would like to thank my Principle Investigator, Dr. Matthew Redinbo, for allowing me to work in his laboratory and mentoring me throughout graduate school. You allowed me to roam about my projects independently without letting me lose track, and allowed me to work on several smaller projects along the way that taught me how to address a scientific question appropriately and how to work as part of a team to complete a project. I would also like to thank you for being completely understanding and respectful during my Father's sickness and passing. I entered graduate school with the idea that life around me would just

stop as I did my studies and I could pick it back up where it left off when I finished, but boy did real life throw me some loops in the past six years. Thank you for being the caring and patient PI that you are. In addition, you have provided an excellent work environment filled with other wonderful people that have made the past six years both incredibly enjoyable and educational. With that in mind, I would like to thank several past members of the Redinbo group, including Dr. Mike Miley, for your guidance early in my graduate career and continued friendship; Dr. Joseph Lomino, who has been a colleague and a friend that helped me move through difficult times both scientifically and personally; Dr. Rebekah Nash, for your caring spirit and inspiring dedication to education; Dr. Yuan Cheng, for mentoring me as a rotation student and filling many days with both knowledge and your humorous antics; Keith Ballentine, for being the ultimate undergraduate then technician and friend, I will always remember our sports conversations, many other past and present members including Major Denise Little, Dr. Sarah Kennedy, Jon Edwards, Bret Wallace, Dr. Krystle McLaughlin, and finally Dr. Michael Johnson, who has been completely there for me as a lab mate, my “lab spouse,” without whom the past several years would not have been nearly as easy or as fun.

I would also like to acknowledge my family. My mother and step-father, who have always supported my continued education; my father, whose passing while I was in graduate school helped me keep perspective; my brother, for the personal support throughout my father’s sickness, that neither of us expected or were able to handle alone, and my in-laws for their continued love and support. I must give a huge thanks to my loving husband, Russ Frazier, for picking up and moving here when I wanted to go to graduate school with no complaints, for being an outlet for all my grumbles and gripes about failed experiments, and loving me and supporting me through the process of graduate school, even when he thought I might never finish. He has encouraged me to pursue my dreams

without accepting anything less, and has been around for the entire process. Finally, I want to acknowledge my daughter, Katie Mabel Frazier, whose arrival completely changed my entire being, helping me realize that my most important job is to be her mother.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Å – Angstrom

α – alpha

AF-2 – activation function-2

Ala, A – alanine

AMBER – assisted model building with energy refinement

Amp – ampicillin

APF – atomic positional fluctuations

ARM – armadillo

Asn, N – asparagine

Arg, R – arginine

Asp, D – aspartic acid

AT – autotransporter

At – *Arabidopsis thaliana*

ATP – adenosine triphosphate

β – beta

C – Celsius

Cα – alpha carbon

CARM – coactivator-associated arginine methyltransferase

CD – circular dichroism

Cl – chloride

CTD- C-terminal domain

CPSF – cleavage and polyadenylation specificity factor

CstF – cleavage stimulation factor

C-terminus, C-term, C-terminal – carboxy terminus

Δ – deletion

DBD – DNA binding domain

DLS – dynamic light scattering

DNA – deoxyribonucleic acid

DNase – deoxyribonuclease

DTT – D,L-dithiothreitol

EDTA – ethylenediaminetetraacetic acid

ERα – estrogren receptor α

F – fluorine

FL – full length

FLC – flowering locus C

fs – femtosecond

g – gravity

GF – gel filtration

Glu, E – glutamic acid

Gly, G – glycine

GST – glutathione S-transferase

GTP – guanosine-5'-triphosphate

H – hydrogen

h – hour

HEAT – huntingtin-elongation-A subunit-TOR

His, H – histidine

HSF – heat shock factor

H4 – histone 4

Iso, I - isoleucine

IPTG – isopropyl β-D-1-thiogalactopyranoside

K, Lys – lysine

K – potassium

K – kelvin

K_d – dissociation constant

kD – kilodalton

λ – wavelength

L – liter

Leu, L – leucine

LB – lurea broth

LBD – ligand binding domain

LIC – ligation independent cloning

Lys, K – lysine

μ – micro

μg – microgram

μL – microliter

μm – micron

μM - micromolar

mdeg – millidegrees

mg – milligram

min – minutes

mL – milliliters

mm – millimeter

mol – mole

mM – millimolar

Met, M - methionine

MBP – maltose binding protein

MD – molecular dynamics

Met, M – methionine

mRNA – messenger ribonucleic acid

MW – molecular weight

Na – sodium

Ni – nickel

nM – nanomolar

nm – nanometer

N-terminus, N-term N-terminal – amino terminus

NLS – nuclear localization signal

NMA – normal mode analysis

NR – nuclear receptor

O – oxygen

OD – optical density

P – phosphate

PAGE - polyacrylamide gel electrophoresis

PD – passenger domain

PDB – Protein Data Bank

PEG – polyethylene glycol

pH – negative log (base 10) of the molar concentration of hydronium ions

Phe, F – phenylalanine

PHYRE – protein homology/analogy recognition engine

PMSF – phenylmethanesulfonyl fluoride

PMEMD – particle mesh ewald molecular dynamics

polyA – polyadenylation

PPARγ – peroxisome proliferator-activated receptor-γ

PRMT – protein arginine methyltransferase

Pro, P – proline

PXR – pregnane X receptor

QHA – quasiharmonic analysis

RIPL – BL21-CodonPlus (DE3)-RIPL cells

RMSD – root mean square deviation

rpm – revolutions per minute

RXR – retinoid X receptor

s – seconds

SAH – S-adenosylhomocysteine

SAM – S-adenosylmethionine

SDS – sodium dodecyl sulfate

Ser, S – serine

SER – surface entropy reduction

SER-CAT – southeast regional collaborative access team

SRC-1 – steroid receptor coactivator-1

TB – terrific broth

Tb – terbium

Tet – tetracyclin

TEV – tobacco etch virus

Thr, T – threonine

T_m – melting temperature

TRIS – 2-Amino-2-hydroxymethyl-propane-1,3-diol

Trp, W – tryptophan

Tyr, Y – tyrosine

UNC-CH – University of North Carolina at Chapel Hill

U.S. – United States

Val, V – valine

WT – wild type

w/v – weight by volume

Yap – *Yersinia pestis* autotransporter protein

% – percent

° – degree

$^{+}$ - positive

$^{-}$ - negative

CHAPTER 1

The *Yersinia pestis* autotransporter YapG contains a fast folding β-helical domain*

1.1 Introduction

1.1.1 Autotransporter Protein Passenger Domains

Autotransporter proteins represent the largest class of secreted proteins from gram-negative bacteria, and are the primary component of type V secretion (*1*). Specialized for protein transport out of the two gram-negative bacterial membranes, autotransporters are made up of an N-terminal signal sequence, a C-terminal β-barrel called the autotransporter domain, and an interior, functional, passenger domain (*2*). The N-terminal signal sequence is used to shuttle the unfolded passenger and autotransporter domains through the Sec translocon of the inner membrane. The C-terminal β-barrel is then folded into the outer membrane of the bacterium, and is utilized as a transport vehicle for the passenger domain peptide out of the bacterium. The shuttling of the passenger domain is not energy dependent (does not require ATP) and requires the passenger domain to remain minimally folded (*3, 4*). The mechanism of this transport is highly debated, as is the degree to which autotransporters facilitate their own transport versus aid from chaperone proteins (*1, 5*).

Once on the exterior of the bacterium, the passenger domain has three potential end locations; they may remain covalently attached to the β-barrel, be cleaved from the β-barrel

*To be submitted for publication

but remain localized to the membrane, or be cleaved from the β-barrel and secreted away from the membrane. These three options are likely associated with the virulence related functions of passenger domains, which includes adhesins, proteases, toxins, and esterases among others (*6*). However, relative to the ubiquitous nature of autotransporter expression in gram-negative bacteria, few passenger domains have been characterized.

1.1.2 Passenger Domains and the β-helical Fold

Despite wide variability in length and sequence, over 97% of passenger domains from autotransporters are predicted to have a right-handed β-helical fold (*7*). The β-helical fold is made up of repeating coils of 3 β-strands connected by variable length loops, with coils repeating on one another to create 3 parallel β-sheets. These coils create a long triangularly shaped domain; each coil has a rise of 4.86 Å (*8*).

The crystal structures of several β-helices have been solved: P.69 pertactin from *Bordetella pertussis* (PDBID: 1dab), hemoglobin protease Hbp (PDBID: 1WXR) and the extracellular serine protease EspP (PDBID: 3SZE), both from pathogenic *Escherichia coli*, the vacuolating toxin p55 VacA from *Helicobacter pylori* (PDBID: 2QV3), and Immunoglobulin A1 Protease from *Haemophilus influenza* (PDBID: 3H09) (*9-13*). The majority of these heretofore solved crystal structures have provided some insight into the significance of the β-helical fold, including the theory that the fold was evolved for transport, not function (*7*). Indeed, several of the passenger domains above include globular domains protruding off the β-helical spine associated with their function as serine protease autotransporters of the Enterobacteriacaea (SPATEs) (*9, 11, 12*). However, pertactin is a much simpler structure made up nearly exclusively with β-spine, with an RGD sequence

motif associated with its function as an adhesin found in one of the loop regions between the β-coils (*10*).

β-helices are particularly interesting from a protein folding perspective because they represent a protein class that has hitherto been understudied by the protein folding community. Not only are β-helices made up of purely β-strand secondary structure (no α-helical content), but they are structural repeat proteins with an associated high contact order and surprisingly low sequence identity despite high structural identity. Only a few β-helical proteins have been biophysically characterized, two of which are passenger domains from autotransporter proteins.

Pertactin and Pet, two passenger domains from *Bordetella pertussis* and *Escherichia coli*, respectively, have both been shown to fold in a three-state process under equilibrium denaturing conditions with the C-terminal end of the passenger domain folding first as a stable core and the N-terminus following (*7, 14*). Pertactin refolding using time-resolved stopped-flow fluorescence has been found to be extremely slow (on the order of hours) and resistant to aggregation (*7, 15*). However, different from results seen in Pertactin and Pet, the P22 tailspike protein and pectate lyase C have been found to undergo a two-state transition from folded to unfolded under equilibrium denaturing conditions (*16-18*). It remains unclear whether the refolding speed or folding pathway of these proteins seen *in vitro* plays any role in the biological function of these proteins.

1.1.3 YapG

Yersinia pestis, the gram negative bacterium responsible for bubonic, pneumonic and septicemic forms of the plague has 9 putative autotransporter proteins that have been found using *in silico* analyses (Yaps C, E, F, G, H, J, K, L, and M) (*19*). The passenger

domains of these autotransporters (Yaps, for *Yersinia pestis* **a**utotransporter **p**roteins) are hypothesized to have virulence associated functions; YapE has been found to be essential for full *Y. pestis* virulence (*20*). The Yaps have passenger domains of variable length (359-3347 amino acids) and sequence (*19*).

YapG, a 994 amino acid autotransporter, is unique from the other Yaps because it is fully secreted (*19*) and it is the one of only two Yaps (G and H) to have repeat regions within the passenger domain. Three repeats (the first two are identical and the third is nearly identical), are found within the C-terminus of the passenger domain (Figure 1.1), and all repeats contain sites within the described recognition sequences for the plasminogen activator Pla, an outer membrane surface protease in *Y. pestis* (*21*). YapG also contains a KDEL sequence motif, which may be associated with retention in the endoplasmic reticulum of targeted cells.

In this work we have overexpressed, refolded, and purified the YapG passenger domain and investigated its fold and structure via equilibrium and time resolved kinetic biophysical studies. We have demonstrated that the secreted passenger domain from YapG is a model β-helix not expected to contain any additional globular domains. Steady-state Circular Dichroism and intrinsic fluorescence measurements in the presence of denaturant indicate a single transition from the native state to the unfolded state, but time-resolved kinetics reveal multi-state folding behavior. Unlike previously studied β-helices, the YapG passenger folds extremely quickly and without formation of aggregates or rate limitation from proline isomerization. The speed of YapG folding contradicts previous notions that passenger domains fold purposefully slow to ensure folding does not occur until full secretion from bacteria (*15*) . Taken together, these results provide the first biophysical analysis of an autotransporter passenger domain in *Y. pestis* and provide new insight into the folding process in β-helical folds.

1.2 Results

1.2.1 Construct Design

The YapG passenger domain extends from the end of the signal sequence on the N-terminus to the start of the C-terminal autotransporter domain (Figure 1.1). SignalP 3.0, a server that predicts signal peptide cleavage sites, predicts the signal sequence to be cleaved between Ala49 and Asn50 (Figure 1.2) (*22, 23*). The end of the passenger domain was predicted using the Pfam protein family database (*24*) by locating the predicted start of the autotransporter domain (residue 718); the C-terminal residue of the YapG passenger is predicted to be at residue 717 (Figure 1.1).

Homology modeling of $YapG_{50-717}$ gave insight into the nature of the passenger domain's tertiary structure (Figure 1.3). The overall fold modeled for the YapG passenger is of a β-helical spine (see Figure 1.3B for a magnification of the β-helical spine), with variable length loops connecting the β-strands of the helix. Beginning at residue 480, the $YapG_{50-717}$ homology model predicts a long region of unstructured sequence that extends to the C-terminus of the passenger domain. In addition, this region (residue 480 onward) of the YapG passenger domain sequence contains a proline-rich region (23 of the passenger domain's 29 proline residues). Proline *cis*/*trans* isomerization is often considered to be the rate limiting step in protein folding (*25*). To overcome these issues, the $YapG_{50-479}$ construct was designed and created using ligation independent cloning. The 50-479 construct should maintain a regular secondary structure and not involve the long unstructured region containing proline rich sequence. Thus, the 50-479 construct was generated with crystallization in mind.

The model's long unstructured region (480-717) contains YapG's repeat region (Figure 1.1), where the surface protease Pla has been shown to cleave YapG from the

surface of *Y. pestis* in at least 3 different places (KR sequences predicted to be cleavage sites are marked in Figure 1.1 and shown in magenta spheres in Figure 1.3A). Mutagenesis of the putative cleavage sites has shown that the first cleavage occurs after residue 512 (Chelsea Lane, Ph.D., Miller lab, data not shown). Thus, it is assumed the functional region of the YapG passenger is within the $YapG_{50-512}$ construct. This construct was also created using ligation independent cloning and was considered the most biologically relevant construct. $YapG_{50-512}$ was used for all studies with the exception of crystallization trials.

1.2.2 The YapG Passenger Domain is β-Helical

Multiple prediction servers were used to evaluate the secondary structure of YapG's passenger. PSIPRED (PSIPRED V3.0) (*26, 27*) predicted the entire sequence to be α-helix free, consisting of only coil and β-strand segments. Likewise, Jpred (*28, 29*) predicts a purely β-strand structure. PredictProtein (www.predictprotein.org) predicts $YapG_{50-512}$ to be 0.0% helix, 48.4% strand, and 51.6% loop (*30*). All servers predicted YapG to be made up of β-strands of varying length (3-16 residues) connected by loops also varying in length (approximately 2-15 residues). This pattern was considered well in line with the canonical makeup of a β-helix (see 1.1.2).

The β-helix prediction program BetaWrap (*31*) was used to compare the YapG passenger domain's raw score with other known β-helices. BetaWrap scores a query sequence against known β-helical structures for the potential of the query to fit into the interacting and stacking residues of those structures. Each score is given a P-value associated with the probability that the same score would be attained using a template from the PDB that is not a β-helix. The $YapG_{50-512}$ construct has a raw score of -22.28 (P-value of 0.0072) for residues 29-157, chosen as the "Best Wrap". This score is in line with other

known β-helices (Pet, Hbp, and pertactin have scores of -21.93, -21.27, and -18.2 with P-values of 0.0048, 1 x 10^{-5}, and 0.0021, respectively) and suggests that YapG's passenger domain has a β-helical fold (*14*).

The YapG constructs were overexpressed and refolded from inclusion bodies after extensive optimization of the refolding protocol. Refolding success was evaluated based on elution from gel filtration at an expected volume in a singular, symmetrical peak, the ratio of soluble protein eluted to soluble aggregate eluted in the void volume, and with a Circular Dichroism (CD) wavelength spectrum.

The CD wavelength spectrum of purified, refolded $YapG_{50-512}$ was used to observe the native secondary structure of the protein. Purified $YapG_{50-512}$ displays a canonical β-helical CD wavelength spectrum (Figure 1.4A, solid downward triangles). The minimum at 215 nm is expected for a protein containing β-strands. No signal is seen for the canonical α-helical minima at 208 nm or 222 nm, indicating predictions that $YapG_{50-512}$ is purely β-helical are correct. The CD spectrum for $YapG_{50-512}$ is very similar to those seen for other passenger domains (*14, 32, 33*). Taken together, the BetaWrap score, secondary structural predictions, and the CD spectrum suggest the $YapG_{50-512}$ construct is a model β-helix, and does not contain any additional globular domains within its passenger.

1.2.3 Thermal Stability of $YapG_{50-512}$

Thermal denaturation of $YapG_{50-512}$ from 10 °C to 90 °C was carried out to determine the melting temperature, T_m, at which the protein becomes 50% unfolded. The minimum CD signal of pure, folded protein (215 nm) was used to monitor the melting process. As the temperature was increased, the signal at 215 nm gradually increased and then went through a sharp transition between 40 °C and 55 °C before plateauing from 60 °C onward. This

signal was converted to percent folded using Eq. (2) and multiplying by a factor of 100, and plotted against the temperature (Figure 1.4B). The T_m was found at approximately 50 °C; $YapG_{50\text{-}512}$ is a thermally stable construct. After melting to 90 °C, the wavelength spectrum of $YapG_{50\text{-}512}$ resembles the canonical random coil signal, confirming that $YapG_{50\text{-}512}$ was completely unfolded during the melting process (Figure 1.4A). Cooling back to 10 °C did not result in recovery of the pre-melting spectrum. Since thermal denaturation of $YapG_{50\text{-}512}$ is not reversible, thermodynamic constants were not calculated.

1.2.4 $YapG_{50\text{-}512}$ Steady-State Unfolding

Equilibrium unfolding of $YapG_{50\text{-}512}$ was performed with urea and monitored by the CD signal at 215 nm. Individual 0.5 μM protein aliquots were denatured in 0.25 M urea steps from 0 M urea to 4 M urea and allowed to equilibrate for a minimum of 1 hour before wavelength spectra were recorded. As the concentration of urea was increased the signal transitioned from a folded, canonical β-helical, signal to an unfolded signal between 1.75 M urea and 3 M urea (Figure 1.5A, selected spectra shown of those collected). Plotted as the percent unfolded versus concentration of urea as described by Greenfield et al. (*34, 35*) in detail and briefed in Methods (Section 1.4), $YapG_{50\text{-}512}$ undergoes a one-step unfolding pathway with transition from folded (F) to unfolded (U) occurring at approximately 2 M urea (Figure 1.5B). Using the transition points between F and U, a plot of ΔG versus the concentration of urea gives ΔG_F, the ΔG of folding, as the y-intercept (Figure 1.5C). Using this method the ΔG_F of $YapG_{50\text{-}512}$ is -5.53 kcal/mol.

In the same manner as the equilibrium unfolding monitored by CD, a measure of the secondary structure throughout the unfolding process, $YapG_{50\text{-}512}$'s intrinsic tryptophan fluorescence in the presence of increasing concentrations of urea was recorded. $YapG_{50\text{-}512}$

has four tryptophan residues within its passenger domain sequence (Figure 1.1: W86, W239, W300, and W427, tryptophans are also colored orange in the YapG passenger homology model, Figure 1.3). Figure 1.6A shows the non-denaturing emission spectrum of $YapG_{50\text{-}512}$ after excitation at 295 nm. As the concentration of urea is increased, the emission spectrum peak shows a decrease in intensity and a slight red shift. The change in intensity of the peak maxima was used to plot the percent folded versus concentration of urea (Figure 1.6B). As observed in equilibrium CD measurements, the transition point from folded to unfolded was observed at approximately 2 M urea and was a one-step transition. The $YapG_{50\text{-}512}$ ΔG_F was observed to be -7.24 kcal/mol (Figure 1.6C).

1.2.5 $YapG_{50\text{-}512}$ Stopped-flow Kinetics of Refolding/Unfolding

Stopped-flow fluorescence experiments were performed to elucidate the kinetics of unfolding/refolding $YapG_{50\text{-}512}$. Unfolding was monitored by mixing 5 μM $YapG_{50\text{-}512}$ in 0 M urea buffer 1:10 with 4 M urea buffer for a final concentration of 0.5 μM $YapG_{50\text{-}512}$ and 3.6 M urea. Here, a signal decrease is associated with the loss of tryptophan fluorescence due to a change in the tryptophan(s) environment (from a folded environment to an unfolded environment). Figure 1.7A shows a 2 second trace of $YapG_{50\text{-}512}$ unfolding fit with a single exponential, 3 parameter fit (Eq. (5)). Residuals for the fit were random (Figure 1.7B), and did not require additional exponential terms. Unfolding of $YapG_{50\text{-}512}$ occurs with a rate constant of 5.61 s^{-1}.

Refolding was monitored by mixing 5 μM $YapG_{50\text{-}512}$ in the presence of 4 M urea 1:9 with 0 M urea buffer (Figure 1.8). Unlike the unfolding process, refolding of $YapG_{50\text{-}512}$ was best fit to a two exponential, 5 parameter fit (Eq. (6)), as seen by the improvement in residuals upon fitting with a double exponential versus a single exponential (Figure 1.8B, C).

Refolding is seen to be a two event process, the first very rapid (< 2 seconds, k_1 = 3.8 s^{-1}) and the second a slower event (< 10 seconds, k_2 = 0.33 s^{-1}). Improvement in the fit was not found by adding additional exponential terms. The multi-state behavior of $YapG_{50-512}$ refolding led to consideration of possible folding mechanisms including possible directionality of folding and whether off pathway aggregation might contribute to wild-type refolding kinetics.

1.2.6 Directionality of Folding

It remains unknown if passenger domains fold directionally, starting the folding process on one terminus and folding sequentially toward the other terminus, or if they fold by overall globular collapse followed by fine adjustments of tertiary structure. The $YapG_{50-512}$ native tryptophan residues (W86, W239, W300 and W427) are spread throughout the sequence, one near each terminus, and two in the middle (Figure 1.1), providing the means to observe folding in different regions of the sequence via single tryptophan mutants. To examine whether the two rates of refolding seen in wild-type $YapG_{50-512}$ are a result of directional folding, tryptophan residues were mutated to phenylalanine, leaving a single tryptophan (3 of the 4 tryptophan residues mutated) on either the N-terminus or the C-terminus (designated WFFF and FFFW) for excitation .

To verify that mutation of 3 of the 4 tryptophans to phenylalanine did not disrupt the stability of $YapG_{50-512}$, CD wavelength spectra and thermal denaturation data were collected (Figure 1.9). The wavelength spectrum of WFFF overlaps with wild-type, and the spectrum for FFFW nearly overlaps with wild-type, suggesting the secondary structure is the same in the mutants as in wild-type (Figure 1.9A). In addition, thermal denaturation indicated that the

mutants were at least as stable as wild-type, with melting temperatures all within a few degrees of wild-type (Figure 1.9B).

Equilibrium fluorescence emission of WFFF and FFFW after excitation at 295 nm was used to observe the individual environments of W86 and W427 upon unfolding with urea. As the concentration of urea was increased, WFFF was immediately quenched with the addition of urea (Figure 1.10A) and FFFW showed no change in emission with the addition of urea (Figure 1.10B). The fluorescence intensity observed for FFFW was approximately 50% of that for WFFF and the peak was shifted to 330 nm versus 355 nm for WFFF. Refolding kinetics of WFFF or FFFW was unable to be recorded due to quenching and the lack of change in emission spectra at W86 and W427, respectively.

Based on the FFFW and WFFF equilibrium and stopped-flow experiments, we hypothesized that the two middle tryptophans make up the majority of the wild-type signal. To test this hypothesis, FWWF was created. Upon denaturation, FWWF demonstrated steady-state fluorescence similar to wild-type in peak maxima shift, but without the overall change in amplitude (Figure 1.11A) To observe refolding of FWWF, a 320 nm cutoff filter was used, which would display nearly the entire spectrum of the unfolded protein and only a fraction of the spectrum for the folded protein (signal decrease upon refolding). FWWF refolds very quickly, within 10 seconds. As seen in wild-type, a double exponential, 5 parameter curve best fits the data, with rates of k_1 = 0.33 s^{-1} and k_2 = 3.07 s^{-1}, nearly identical to wild-type (Figure 1.11B). Additional exponential terms did not increase the quality of the fit or the randomness of the residuals (Figure 1.11C, D). Thus, the overall change in environment (folding process) observed in wild-type $YapG_{50\text{-}512}$ (WWWW) is the same as that seen with FWWF.

1.2.7 Potential for Aggregation during Refolding

Off pathway aggregation was considered as a possible source of the multi-state protein folding seen in $YapG_{50-512}$. To investigate this possibility, pure $YapG_{50-512}$ and refolded $YapG_{50-512}$ were run over a gel filtration column. Pure $YapG_{50-512}$ eluted as a single, symmetrical peak, while refolded $YapG_{50-512}$ eluted in multiple peaks (Figure 1.12). The two peaks seen for $YapG_{50-512}$ straddle the pure $YapG_{50-512}$ peak. The two peaks are connected and elute with the second peak dominant to the first and the first appearing as a shoulder on the second peak. Due to the closeness of their elution volume, it is likely that the two species are the same size, but are two differently folded populations of $YapG_{50-512}$. There was no peak associated with the void volume of the gel filtration column (approximately 45 mLs); no aggregation occurs during the refolding process. Concentration limitations prevented collection of these peaks and analysis using DLS.

1.2.8 Crystallization Trials

The $YapG_{50-479}$ construct, which has a more definitive end in its secondary structural prediction versus $YapG_{50-512}$, was designed primarily for crystallization considerations. Pre-screening of $YapG_{50-479}$ was done at multiple concentrations (2mg/mL – 10 mg/mL). Initial crystal hits of $YapG_{50-479}$ were found using the Rigaku high-throughput crystallization screening robot at the UNC Biomolecular X-ray Crystallography Facility. Screens utilized included pH Clear, Classics Lite, PEGs, and PEGs II (QIAGEN) among others. Initials hits found all came from the PEGs screens, and all included 0.2 M lithium sulfate and 0.1 M Tris pH 8.5, with varying high percentage PEGS: 30% PEG 3000, 30% PEG 4000, or 25% PEG5000. All conditions produced needle-like rods, often growing in stacks. Initial hits were optimized using several methods: optimization screens, micro-seeding and macro-seeding,

decoupling, additive screens, volume/ratio screens (total volume of drop as well as the ratio of protein to mother liquor).

Crystals of $YapG_{50-479}$ (Figure 1.13) were found to diffract poorly or not at all. Multiple approaches were utilized to encourage YapG crystals to diffract. The protein concentration for crystallization was reduced to eliminate aggregation during concentration and crystallization after Dynamic Light Scattering (DLS) indicated that $YapG_{50-479}$ aggregates at concentrations beyond 2.5 mg/mL. Trays were subsequently all set up at 2.5 mg/mL, at which DLS indicated a single, monodisperse species (polydispersity < 10%).

Crystal trays were set up at 4 °C (versus 20 °C) to reduce the likelihood of any thermal denaturation during crystallization. Cleavage of the His tag used for purification was performed, and found to make a slight difference in crystal morphology but not in diffraction quality.

Surface entropy reduction (SER) mutations were introduced into $YapG_{50-479}$ to exchange clusters of high entropy amino acids (lysine, glutamic acid, for example), predicted to be on the surface of the protein where crystal packing occurs, for alanine. Clusters of high entropy amino acids were chosen using the Surface Entropy Reduction Prediction Server (*36*). Three clusters were identified in YapG, numbered SER1-SER3 based on their score from the server. Clusters were identified in the YapG homology model and found to be reasonably placed with respect to the protein surface (Figure 1.14A). The higher the score, the more likely the cluster's mutagenesis to alanine was predicted to contribute to a change in crystal packing and potentially diffraction quality. The SER2 cluster (KKQ→AAA) was on the N-terminus of the wild-type construct and when introduced to the sequence, the construct was truncated to begin with the SER2 mutation. SER1 (EKK→AAA), in the middle of $YapG_{50-479}$, was also cloned with the SER2 start and was therefore termed SER1/2. SER3 (EK→AA) was cloned into the original 50-479 construct.

All SER mutations (SER1/2, SER2, and SER3) were overexpressed, refolded, and purified. SER mutations were considered properly refolded by comparison to wild-type CD wavelength scans (Figure 1.14B). SER mutations were not found to have any new crystallization conditions and did not prove beneficial to crystallization of $YapG_{50-479}$.

1.3 Discussion and Future Directions

1.3.1 The $YapG_{50-512}$ Passenger is a Fast Folding β-helix

YapG, a secreted autotransporter from *Y. pestis*, is unique because of its redundant cleavage region and KDEL sequence (Figure 1.1) (*19*). The homology model built for the YapG passenger is strikingly similar to the P.69 pertactin structure (Figure 1.3) (*10*), and while possibly correct, highlights the weakness of homology models without many structural templates. However, the current homology model is useful for development of hypotheses in the absence of an X-ray crystal structure. Homology modeling combined with multiple secondary structural predictions of the YapG passenger domain predict the passenger is made up of a β-helical spine beginning with the N-terminus that extends to residue 479, followed by an extended region of little to no structural content, including the repeat region, to the beginning of the autotransporter domain. The functional region of the YapG passenger is presumed to be contained within the minimum cleaved fragment ($YapG_{50-512}$, Chelsea Lane, Ph.D., Miller lab, data not shown).

$YapG_{50-512}$ has been found to be entirely β-strand in nature (Figure 1.4A), with no apparent α-helical content. CD spectra indicate a canonical β-helical signal, with minima around 215 nm. The 50-512 construct was found to be thermally stable, with a T_m of 50 °C. CD and intrinsic tryptophan fluorescence equilibrium studies of $YapG_{50-512}$ display a single step, two-state folding (Figures 1.5 and 1.6). The ΔG_U calculated from these experiments is

within that expected for stable proteins, -3 to -15 kcal/mol (*37*). Time-resolved kinetics using stopped-flow fluorescence, however, suggest a multi-state folding process.

The most striking result from $YapG_{50-512}$ refolding kinetics studies was the speed of refolding. $YapG_{50-512}$ was observed to completely refold *in vitro* within 10 seconds (Figure 1.9) and without formation of aggregates (Figure 1.13). This result makes $YapG_{50-512}$ the fastest refolding β-helix to date, and 3 orders of magnitude faster than *in vitro* refolding of pertactin (*15*) and more in line with the predicted rate of refolding based on contact order (*38*). Interestingly, despite the disparity between their *in vitro* refolding rates, neither $YapG_{50-512}$ nor pertactin are observed to form aggregates during refolding. However, like pertactin, $YapG_{50-512}$ exhibits multi-state folding, leading the role of each folding step to be questioned. Gel filtration of refolded $YapG_{50-512}$ suggests that two same size, but slightly different shaped protein forms exist in solution (Figure 1.12). It is possible that two differently folded species form during refolding, and the folding of each of these two species makes up the multi-state behavior seen in *in-vitro* refolding kinetics.

Further investigation into the multi-state behavior of YapG refolding includes refolding and unfolding $YapG_{50-512}$ to a variety of final denaturant concentrations to create a Chevron plot. If the Chevron plot displays non-linearity it will act as further validation that $YapG_{50-512}$ exhibits multi-state refolding. In addition, insight into the mechanism of refolding remains a goal of this research, with the desire to define the folding events associated with each of the two defined refolding rates. Current progress toward elucidating the refolding mechanism is discussed below.

1.3.2 Role of Proline Isomerization

For other β-helices, namely the P22 tailspike and pectate lyase C, proline *cis*/*trans* isomerization has been found to be associated with a second, slower rate constant during refolding experiments (*16, 32*) and in many instances as the rate-limiting step in protein folding (*39*). In this process, the isomerization between the *cis* and *trans* form of this amino acid's side chain slow the folding rate of the protein. The rate of prolyl peptide isomerization in unstructured peptides is ($0.01\text{-}0.1 s^{-1}$) (*40, 41*).

The $YapG_{50\text{-}512}$ construct contains eight prolines; without the crystal structure it is not possible to know whether any *cis* prolines exist in the native structure. However, $YapG_{50\text{-}512}$ refolding (Figure 1.9, $k_1 = 3.8\ s^{-1}$ and $k_2 = 0.33\ s^{-1}$) is too fast for proline isomerization to be the source of either of the rates found in refolding experiments.

1.3.3 Directionality of Folding

Previous β-helical studies have suggested that some autotransporter passenger domains have a stable core in the C-terminus which folds before the N-terminus (*7, 14*). Selective mutation of three of the four tryptophan residues in $YapG_{50\text{-}512}$ to phenylalanine allowed for single, C-terminal (W427) or N-terminal (W86), intrinsic tryptophan fluorescence. Using this experimental set up, it was expected to be possible to extrapolate whether the folding of a stable core on one terminus might represent one of the two refolding rates seen in $YapG_{50\text{-}512}$ refolding. Thus, if the YapG passenger contains a C-terminal stable core as seen in Pet and pertactin, multi-state refolding would be observed using FFFW and two-state refolding would be observed with WFFF. CD and thermal denaturation indicated that both WFFF and FFFW maintained their secondary structure and stability (Figure 1.10). However, steady-state emission spectra after excitation at 295 nm indicated that only the N-

terminal tryptophan experienced an environmental change upon denaturation with 4M urea (Figure 1.11) and this change was very slight. Refolding kinetics of WFFF or FFFW was unable to be recorded due to the lack of change in emission spectra at W86 or W427.

The lack of change in emission maxima and intensity for W86 and W427 suggested the environmental change observed by the two internal tryptophans, W239 and W300, upon unfolding/refolding are responsible for wild-type emission behavior. Indeed, FWWF was observed to undergo a red shift in its emission maxima upon denaturation and its refolding kinetics was nearly identical to wild-type. The direction of the refolding was opposite that observed in wild-type because FWWF did not undergo a significant intensity decrease upon denaturation and a different cutoff filter was used to observe the refolding event (320 nm cutoff versus 305 nm), but the fitted rates are nearly identical (FWWF k_1 is 19% less than WWWW, k_2 is identical) to those observed in wild-type.

An additional method of elucidating whether $YapG_{50-512}$ folds directionally would be to attach a fluorescein molecule on each terminus of the protein, creating an N-terminally and a C-terminally fluorescent construct, and measuring their respective refolding kinetics as measured for pertactin by Junker and Clark (*15*). This method would best be designed with a crystal structure of YapG so that surface exposed regions could be selected for placement of the fluorescein molecules (as done for pertactin). However, despite limitations from the lack of a crystal structure for YapG, the fluorescein experiment would be ideal for passenger domains like YapG because they do not contain native cysteine residues (thought to be associated with the need for passenger domains to remain unfolded until secretion from the bacterial outer membrane) and the introduction of a single cysteine would guarantee that only one fluorescein molecule would specifically bind to the designed constructs. In the case of pertactin, site specific fluoresceine labeling as described above showed that the N-terminally and C-terminally labeled proteins both refolded at the same rate *in vitro*,

suggesting the passenger domain's termini fold on similar time scales, rather than one terminus folding preferentially before the other (*15*).

1.3.4 Crystallization of Yaps

The crystal structure of the YapG passenger domain was sought for use in determining its structure-function relationship in combination with functional data from the Miller lab at UNC Chapel Hill. The $YapG_{50\text{-}479}$ construct was designed with crystallization in mind by stopping the construct before the predicted long, unstructured repeat/cleavage region. Crystals were grown in multiple PEG containing conditions but none were found to diffract well enough for structure determination using either in-house or synchrotron beam sources.

Future attempts at crystallization could include trials of the $YapG_{50\text{-}512}$ construct. The inclusion of a portion of the unstructured region is not likely to improve crystallization, but DLS results suggest that at 2.5 mg/mL $YapG_{50\text{-}512}$ is a stable dimer instead of a combination of monomer and dimer in solution. A more stable oligomer may improve crystal packing and diffraction. In addition, crystallization after the addition of Maltose Binding Protein to one terminus of the YapG passenger domain construct, or another small, soluble protein which easily crystallizes could be beneficial to crystallization.

1.4 Methods

Homology Modeling

The $YapG_{50-717}$ homology model was built using the HHpred homology detection and structure prediction using the HMM-HMM comparison online server (*42*). The model was built using multiple templates chosen by the server; MODELLER (*43*) was used to build the model.

Cloning, Expression, Refolding and Purification

DNA for full length YapG was generously donated by Dr. Virginia Miller. $YapG_{50-512}$ was incorporated into the LIC (ligation-independent cloning) vector pMCSG7, which includes a hexi-His tag N-terminal to $YapG_{50-512}$ connected by a tobacco etch virus (TEV) protease cleavage site (*44*). $YapG_{50-512}$ was transformed into *E. coli* BL21-CondonPlus(DE3)-RIPL competent cells (Stratagene) and grown in Luria broth (LB) supplemented with 50 µg/ml ampicillin, chloramphenicol, streptomycin, and 25 µg/ml tetracycline each at 37°C with shaking. Cultures were started from a single colony in 5 mL over 8 hours and used to inoculate a 100 mL overnight culture. The resultant pellet was used to inoculate 6 1 L cultures in the presence of above antibiotics and 40 µL antifoam (Sigma-Aldrich) per liter LB. Cells were grown at 37°C until OD_{600} reached 0.9-1.0. Protein expression was induced using a 1 mM final concentration of isopropyl β-D-1-thiogalactopyranoside (IPTG) and grown for an additional 2 hours at 37°C. Cells were centrifuged at 4,500 x g for 15 minutes at 4°C and stored at -80°C.

Individual pellets were thawed for YapG inclusion body isolation, refolding and purification. After thawing for several minutes on ice, a pre-mixed solution of lysis buffer (50 mM Tris pH 8.0, 200 mM NaCl) plus 1 protease inhibitor tablet (Roche), a pinch of lysozyme (MP Biomedicals), 350 µL Triton-X 100 (Alfa-Aesar), and 3 µL benzonase nuclease (Sigma-

Aldrich) were added to the pellet. After incubating on ice for several minutes, the pellet was resuspended into the pre-mixed lysis solution and then sonnicated using a Branson sonic dismembrator at 40% amplitude for 3 minutes using pulses of 0.5 sec with an off time of 1.0 sec between pulses, and spun at 17,000 rev/min for 1 hour. After centrifugation, the resultant pellet was washed in lysis buffer containing 0.5% Triton-X 100 2 times; the first wash included repeating the sonication procedure after resuspension and the remaining washes included resuspension with a dounce homogenizer. Two additional washes were performed in the absence of Triton-X 100. All washes were followed with 10-12 minute spins; after each wash the supernatant was discarded. The resultant inclusion body pellet was solubilized in 8 M urea overnight at 4 °C on a rotary shaker.

Solubilized inclusion bodies were centrifuged at 5,000 x g for 10 min to remove insoluble material, diluted to 1.0 mg/mL in 8 M Urea and dialyzed against 2 L refolding buffer (50 mM potassium phosphate pH 8.0, 350 mM NaCl) plus 4 M urea for 8-24 hours, then repeated step wise with buffer containing 2 M urea, 1 M urea, 0.5 M urea, 0.1 M urea, and lastly buffer A (20 mM potassium phosphate pH 7.4, 500 mM NaCl, 50 mM Imidazole, 20 mM Urea, 0.02% NaN_3). After dialysis into buffer A, refolded protein was passed through a 0.2 µm filter (Millipore) and loaded onto a HisTrap crude column (GE Healthcare) equilibrated with buffer A. YapG was eluted with buffer B (20 mM potassium phosphate pH 7.4, 500 mM NaCl, 500 mM Imidazole, 20 mM Gdn-HCl, 0.02% NaN_3), pooled and loaded onto a 16/60 desalting column (GE Healthcare) equilibrated with buffer C (50 mM Tris pH 8.8, 150 mM NaCl, 5% glycerol). Protein fractions were collected and concentration estimated using A_{280}. TEV protease was added to pooled protein fractions at 3% mass and left to dialyze overnight against buffer D (50 mM Tris pH 8.0, 150 mM NaCl, 1 mM DTT, 5% glycerol) at 4°C. After overnight dialysis, TEV cleaved protein was passed over a crude HisTrap column equilibrated with buffer C and collected from the flow-through. Protein was

concentrated to minimal volume and passed over a Superdex 200 column (GE Healthcare) equilibrated in buffer C. Protein fractions were collected from S200 elution, concentrated to approximately 2.5 mg/mL and frozen in 40 µL aliquots at -80°C for storage.

Refolding protocol was optimized after trials of different refolding methods, including rapid dilution into refolding buffer where the solubilized inclusion bodies in denaturant (8 M urea or 6 M Gdn-HCl) were introduced drop-wise at a set rate using a syringe pump into a large volume of 0 M denaturant refolding buffer with rapid stirring (final concentration of approximately 0.2 mg/mL protein). Setup allowed for drops to be introduced into the 0 M denaturant buffer directly next to the stir-bar to allow for the most rapid dilution out of denaturant possible. This method was attempted with multiple drop rates and final protein concentrations with slower rates (0.5 mL/min) and lower final protein concentrations (0.2 mg/mL) producing a higher refolding efficiency. The final dilute solution (approximately 1 L, depending on the starting concentration of inclusion bodies) was then dialyzed overnight against 8 L of 0 M denaturant refolding buffer, for a final urea or Gdn-HCl concentration of at least 20 mM (concentrations lower than 20 mM caused protein to crash out). In comparison to step-wise dialysis refolding, the rapid dilution protocol was less efficient (Table 1.1) and its use was discontinued after the step-wise dialysis method was optimized for a higher yield.

Circular Dichroism

All circular dichroism experiments were performed on an Applied Photophysics Chirascan CD spectrometer (Applied Photophysics, Leatherhead, Surrey, U.K.) in CD buffer (10 mM potassium phosphate pH 8.0, 200 mM KF). Individual protein aliquots were incubated with urea ranging from 0 M to 4 M in 0.25 M steps. After incubation for a minimum of 1 hr, CD spectra were collected of each protein/denaturant mixture. Wavelength scans

were recorded at 10°C from 260 to 200 nm in quartz cuvettes (Hellma) with a path length of 0.1 cm with 50,000 points collected at each 0.5 nm step. CD signal in ellipticity was converted to mean residue ellipticity (MRE) using Eq. (1), where θ is ellipticity in millidegrees, M_r is the sample molecular weight divided by the number of amino acids, *C* is the sample concentration in grams per liter, and *l* is the path length in cm.

$$[\theta] = [(\theta \times M_r \times 0.1)/(C \times l)] \quad \text{Eq. (1)}$$

The change in free energy, ΔG, of denaturation was calculated as described by Greenfield (*34*) by first calculating the fraction folded, F_i, at each denaturant concentration using Eq. (2):

$$F_i = \left[{([\theta]_{obs} - [\theta]_U)}\Big/{([\theta]_F - [\theta]_U)} \right] \quad \text{Eq. (2)}$$

where $[\theta]_{obs}$ is the MRE at a given denaturant concentration, and $[\theta]_F$ and $[\theta]_U$ are the MRE when the protein sample is completely folded and unfolded, respectively. The fraction folded is simply converted to the folding constant, K_F.

$$K_F = {F_i}\Big/{(1 - F_i)} \quad \text{Eq. (3)}$$

The change in free energy of folding, ΔG_F, is then calculated using Eq. (4), where R is the gas constant (1.98 cal/mol) and T is the absolute temperature In Kelvin.

$$\Delta G_F = -RTln\, K_F \quad \text{Eq. (4)}$$

Thermal denaturation was observed by monitoring the CD signal at 215 nm over heating from 10 °C to 90 °C using a 1 °C step with a tolerance of 0.2 °C and a 30 second hold at each temperature. The CD signal for a given protein sample was converted to percent folded at each temperature by first converting to the fraction folded, *F*, using Eq. (2)

where $[\theta]_{obs}$ is the CD signal at temperature, T, and $[\theta]_F$ and $[\theta]_U$ are as describe above. The fraction folded was then converted to percent folded by simply multiplying F by a factor of 100. The temperature at which the protein sample was 50% unfolded is then set at the T_m of the sample, and considered a measure of protein stability.

Intrinsic Tryptophan Fluorescence

Steady-state fluorescence scans were performed on a SPEX Fluorolog-3 Research T-format Spectrofluorometer at 20 °C in 50 mM $K_xH_yPO_4$ pH 8.0, 150 mM NaCl. Emission scans were collected from 300-400 nm after excitation at 295 nm. Slit widths were set at 1 mm for excitation and 5 mm for emission.

Kinetics of refolding/unfolding was measured on a stopped-flow fluorimeter (SX.18MV, Applied Photophysics) in the Almeida lab (University of North Carolina at Wilmington Department of Chemistry and Biochemistry). The fluorescence signal recorded was the intrinsic tryptophan emission after excitation at 295 nm. After mixing 9:1 within the stopped-flow the concentration of urea was 0.4 M for refolding and 3.6 M for unfolding, and the protein concentration was 0.5 μM. Unfolding kinetics were fit with Eq.(5), a single exponential, 3 parameter fit.

$$y = y_0 + ae^{-bx} \qquad \text{Eq. (5)}$$

Refolding kinetics were fit with both Eq. (5) and Eq. (6), a double exponential, 5 parameter fit. Residuals were compared to determine the best fit in all refolding experiments.

$$y = y_0 + ae^{-bx} + ce^{-dx} \qquad \text{Eq. (6)}$$

Dynamic Light Scattering

Dynamic light scattering experiments were performed on a Wyatt DynaPro Dynamic Light Scattering Plate Reader in 50 mM Tris pH 8.8, 150 mM NaCl, 5% glycerol at room temperature (23-25 °C).

1.5 Figure Legends

Figure 1.1 Schematic of YapG. The YapG full length sequence is made up of a signal sequence (ss, pink), an autotransporter domain (green), and a passenger domain (blue). The region of the passenger predicted to contain little to no secondary structural elements is shown with a dashed line. Constructs used in this study (50-479 and 50-512) are indicated with a bar and circled residue numbers. The repeat region within the passenger domain (residues 509-640) is shown in orange, with cleavage sites marked (Chelsea Lane, Ph.D., Miller lab, data not shown). Locations of tryptophan residues are also marked with Ws and their residue number.

Figure 1.2 SignalP output for YapG. The SignalP gram-negative neural network predicts YapG to have a signal peptide that is cleaved between Ala49 and Asn50. The S-score is associated with the likelihood of a sequence to be involved in a signal peptide. The C-score, or cleavage site score, should only increase at the cleavage site. The Y-score is the derivative of the C-score combined with the S-score. Generated using SignalP (*22, 45*).

Figure 1.3 Homology model of YapG passenger domain. A. Model of the entire passenger domain, with cleavage sites for the Pla surface protease shown in magenta spheres and the remainder of the passenger domain in cyan ribbon. Locations of tryptophan residues are shown in orange sticks. **B.** Close up of the 50-479 truncated construct to optimize for sequence predicted to contain secondary structure.

Figure 1.4 Thermal denaturation of $YapG_{50-512}$. A. Wavelength spectra of $YapG_{50-512}$ after purification, before melting, at 20 °C (black, downward triangles) and after thermal denaturation to 90 °C (grey circles). **B.** Percent unfolded versus temperature of $YapG_{50-512}$. The T_m of $YapG_{50-512}$ is shown as the temperature at which the $YapG_{50-512}$ is 50% unfolded, approximately 50 °C.

Figure 1.5 Equilibrium denaturation of $YapG_{50-512}$ monitored with circular dichroism. A. CD wavelength spectra of $YapG_{50-512}$ in the presence of selected concentrations of denaturant. As the concentration of denaturant (urea) increases, the signal at 215 nm associated with β-strand content is lost in favor or a more random coil signal. **B.** $YapG_{50-512}$ percent unfolded versus concentration of urea goes through a single transition, indicating two-state folding behavior. **C.** Change in free energy, ΔG (kcal/mol), versus concentration of urea for the transition region of $YapG_{50-512}$ equilibrium unfolding fit with linear regression and extrapolated to the axes to determine the ΔG of folding. Points from **B** used in **C** are color matched for clarity.

Figure 1.6 Equilibrium denaturation of YapG$_{50-512}$ monitored with intrinsic tryptophan fluorescence. A. Equilibrium intrinsic tryptophan fluorescence emission spectra of YapG$_{50-512}$ in the presence of varying concentrations of urea after excitation at 295 nm. **B.** Percent unfolded YapG$_{50-512}$ at 350 nm versus concentration of urea. The transition from folded to unfolded is colored to indicate the points used in **C.** ΔG in kcal/mol versus concentration of urea. Transition fit to a linear regression and extrapolated to the axes to determine the ΔG of folding.

Figure 1.7 Stopped-flow unfolding of YapG$_{50-512}$. YapG *in vitro* unfolding is very fast and single exponential. **A.** Unfolding kinetics shown as a function of intrinsic tryptophan fluorescence emission after excitation at 295 nm. Experimental data (open circles) fit with a single exponential function (black line). **B.** Random residuals are shown for data fit to a single exponential function.

Figure 1.8 Stopped-flow refolding of YapG$_{50-512}$. YapG$_{50-512}$ is the fastest *in vitro* refolding passenger domain to date. **A.** Refolding kinetics is shown as a function of intrinsic tryptophan fluorescence emission after excitation at 295 nm. Experimental data (open circles) are fit with a single exponential fit (red line) and a double exponential fit (black line) **B.** Random residuals are shown for data fit to a double exponential function. **C.** Non-random residuals are shown for data fit to a single exponential function.

Figure 1.9 Comparison of W mutants to wild-type YapG$_{50-512}$. Mutation to exclude 3 of the 4 tryptophans does not interfere with the overall secondary structure (**A**) or the thermal stability (**B**) of YapG$_{50-512}$.

Figure 1.10 Intrinsic fluorescence of WFFF and FFFW. Steady-state intrinsic tryptophan fluorescence emission after excitation at 295 nm of WFFF (**A**) and FFFW (**B**) in non-denaturing conditions (black) and in the presence of 4 M urea (red).

Figure 1.11 FWWF monitors the same process as wild-type. A. Steady-state intrinsic tryptophan fluorescence emission after excitation at 295 nm of FWWF in non-denaturing conditions (black) and in the presence of 4 M urea (red). **B.** Kinetics of refolding is shown as a function of intrinsic tryptophan fluorescence emission. Experimental data (open circles) is shown fit to a single exponential fit (red line) and a double exponential fit (black line). Random residuals are plotted for the double exponential fit (**C**) and non-random residuals are plotted for the single exponential fit (**D**).

Figure 1.12 Refolded YapG$_{50-512}$ does not induce aggregation. Overlaid elution profiles from gel filtration of purified protein (solid line) with purified protein that was first denatured in 6 M urea and then refolded by quick dilution to 0.6 M urea (dashed line) to mimic conditions inside the stopped-flow cuvette. No peak is observed in the void volume of the gel filtration column where soluble aggregate would be expected to elute. A peak with a shoulder of the refolded sample may indicate two differently folded species, likely of the same size but of slightly different shapes.

Figure 1.13 Crystals of YapG$_{50-479}$. Best crystals grown during YapG$_{50-479}$ crystallization optimization. Morphologies included rods of approximately 10 x 10 x 200 microns (left), plates, and approximately 10 x 40 x 80 trapezoidal shaped (right) crystals. Conditions detailed in Methods.

Figure 1.14 Surface entropy reduction mutations introduced into YapG$_{50-479}$. A. Clusters of amino acids (shown on the YapG$_{50-479}$ homology model) chosen for mutation to alanine in order to reduce surface entropy and potentially aid in crystallization of YapG$_{50-479}$ are boxed with their respective scores from the Surface Entropy Reduction Prediction Server. **B.** CD wavelength spectra of SER mutations versus the wild-type YapG$_{50-479}$ signal. Mutations were not found to affect the nature of YapG's secondary structure.

Table 1.1 Refolding Protocol Comparison

	Refolding Protocol	
	Rapid Dilution	Step-wise Dialysis
Starting culture volume	1.5 L	1.5 L
Starting mass inclusion body	200 mg	200 mg
Days refolding requires	1	> 3
Yield after affinity	5-8 mg	30-40 mg
Yield after final purification	0.5-3 mg	4-15 mg

Figure 1.1

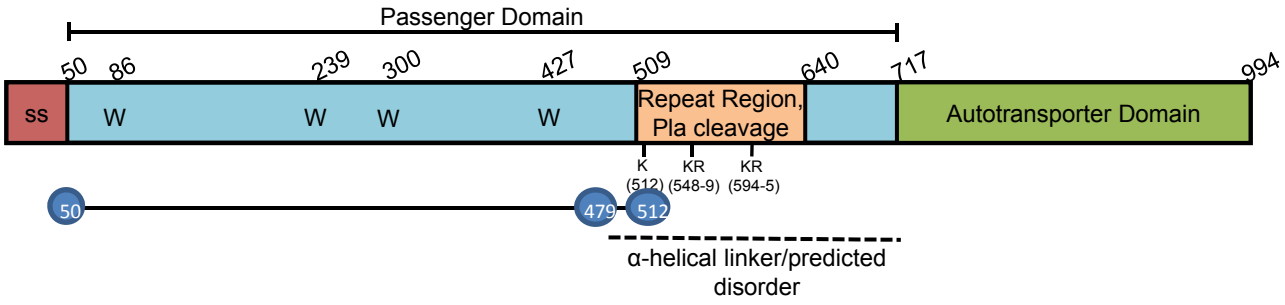

Passenger Domain
50
86
239
300
427
509
640
717
994
ss
W
W
W
W
Repeat Region,
Pla cleavage
Autotransporter Domain
K
(512)
KR
(548-9)
KR
(594-5)
50
479
512
α-helical linker/predicted
disorder

Figure 1.2

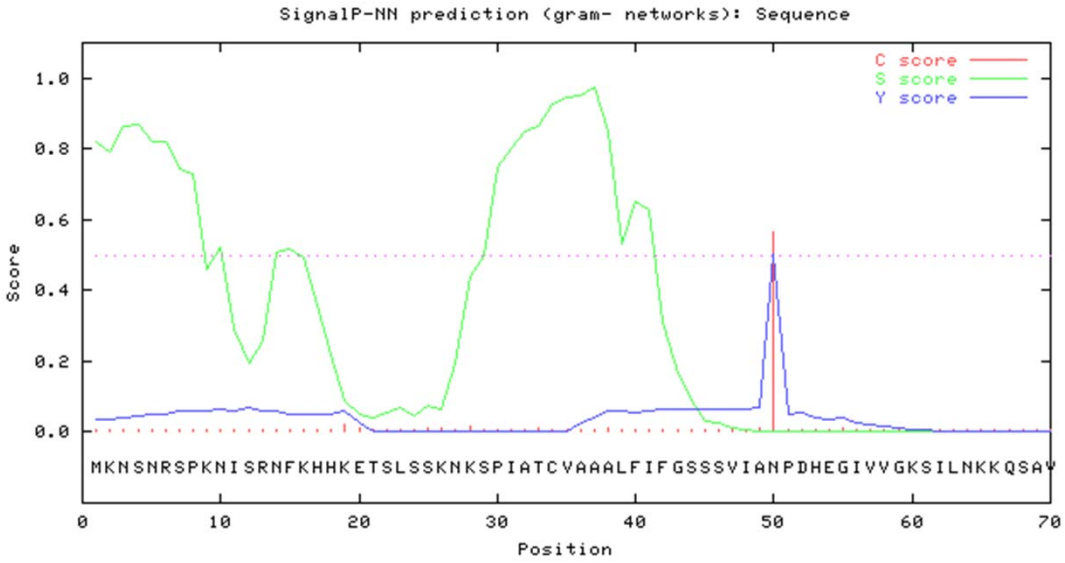
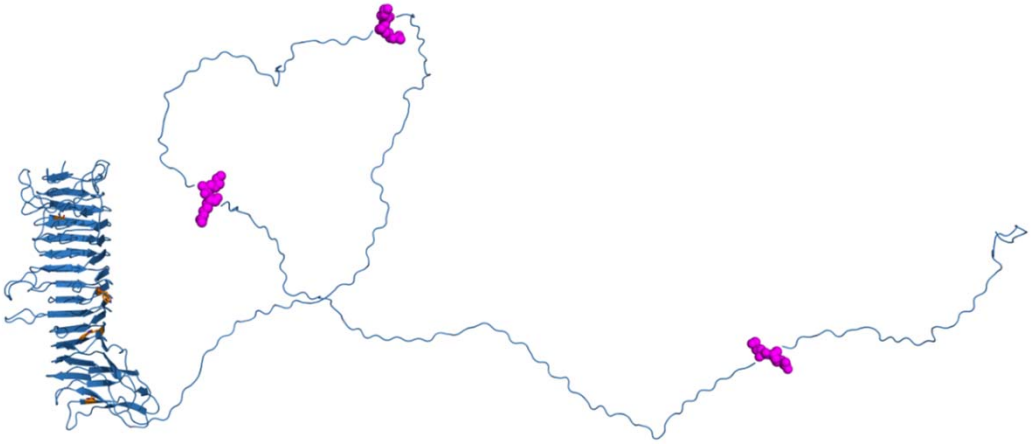

SignalP-NN prediction (gram- networks): Sequence
C score
S score
Y score
Score
1.0
0.8
0.6
0.4
0.2
0.0
MKNSNRSPKNISRNFKHHKETSLSSKNKSPIATCVAAALFIFGSSSVIANPDHEGIVVGKSILNKKQSAV
0
10
20
30
40
50
60
70
Position

Figure 1.3

A



B

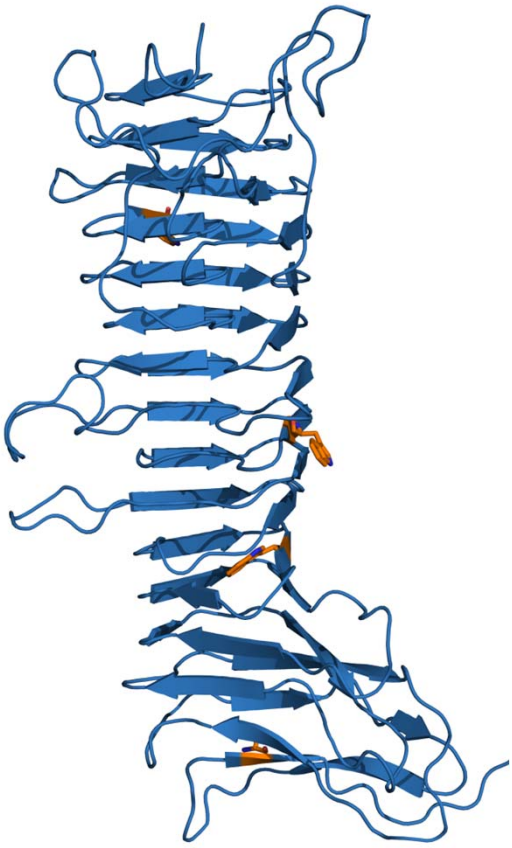
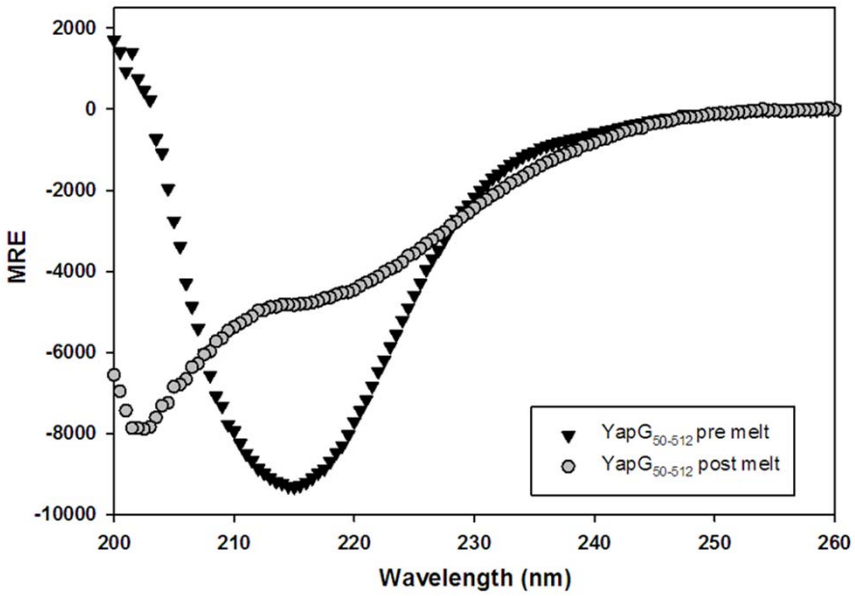


Figure 1.4

A


2000
0
-2000
-4000
-6000
-8000
-10000
MRE
200
210
220
230
240
250
260
Wavelength (nm)
YapG50-512 pre melt
YapG50-512 post melt

B

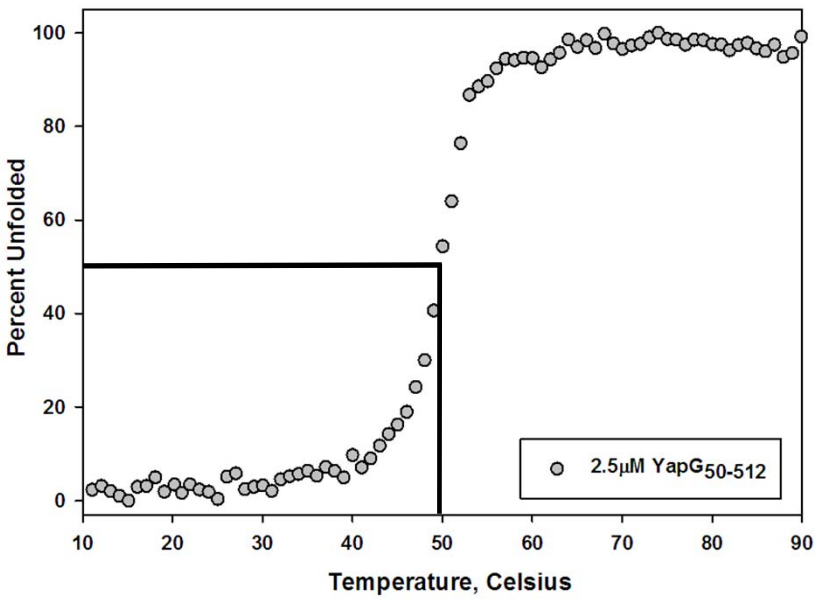

100
80
60
40
20
0
Percent Unfolded
10
20
30
40
50
60
70
80
90
Temperature, Celsius
2.5µM YapG50-512

Figure 1.5

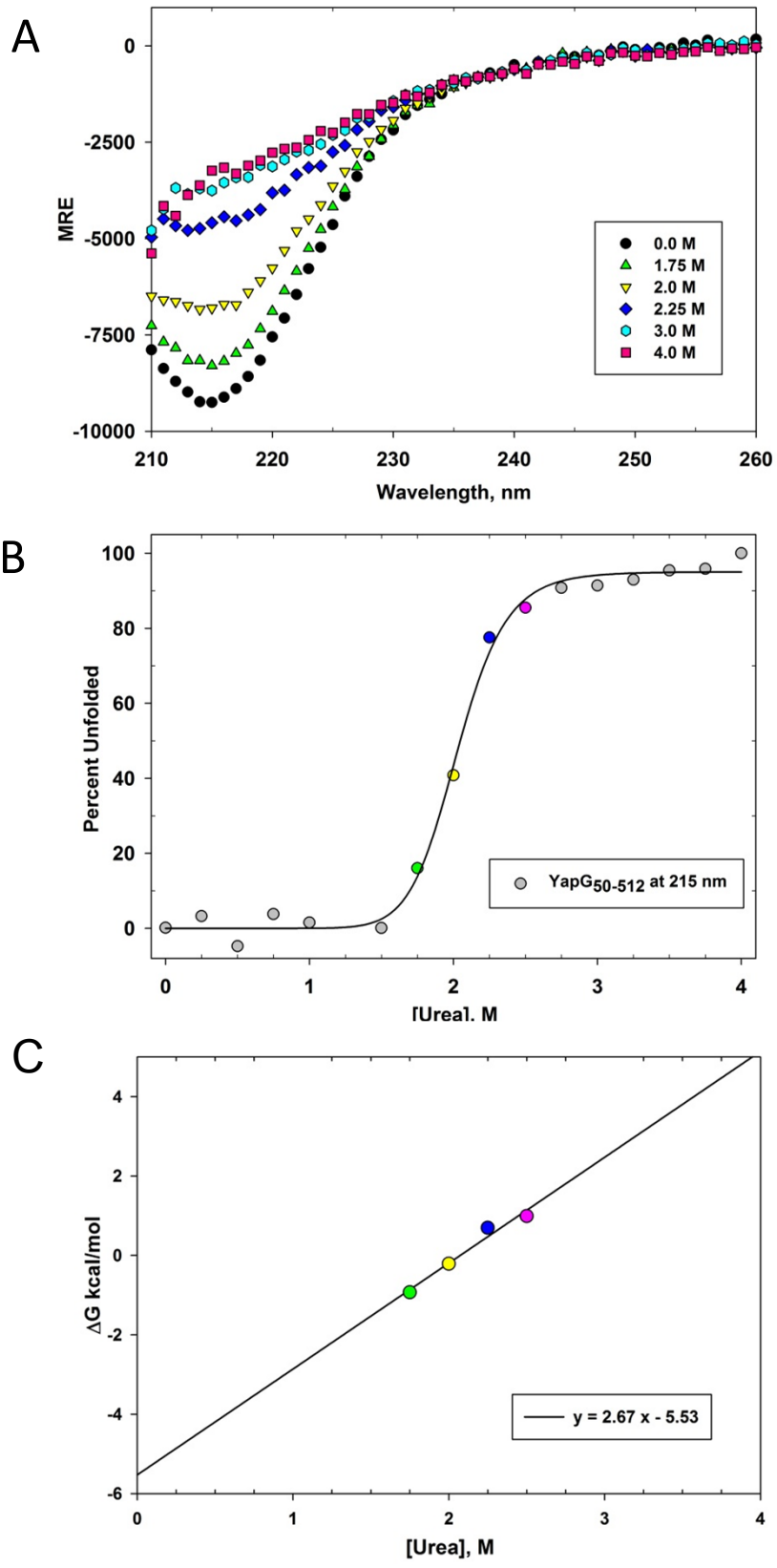
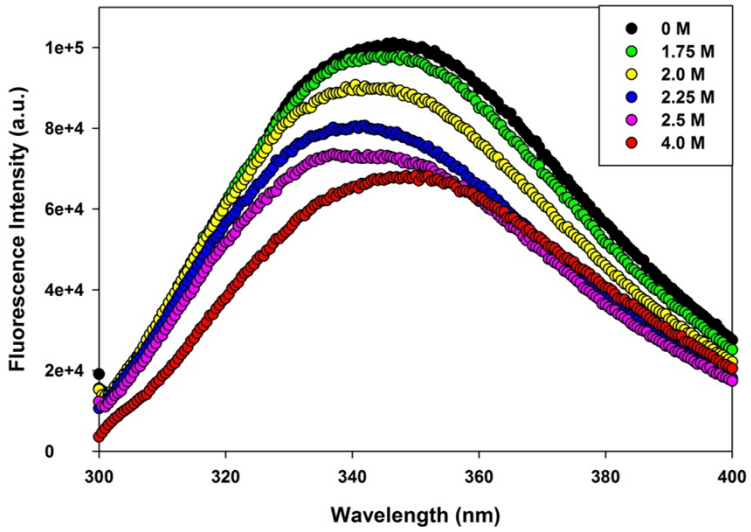
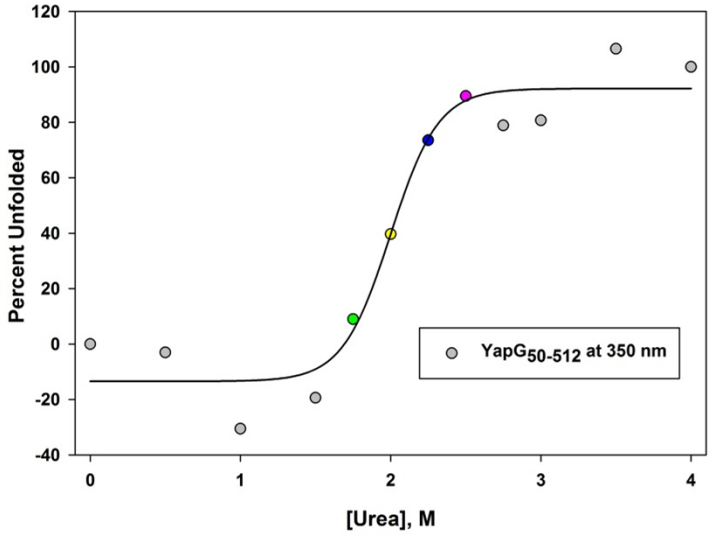

A
MRE
Wavelength, nm
0.0 M
1.75 M
2.0 M
2.25 M
3.0 M
4.0 M
B
Percent Unfolded
[Urea], M
YapG50-512 at 215 nm
C
ΔG kcal/mol
[Urea], M
y = 2.67 x - 5.53

Figure 1.6

A



B



C

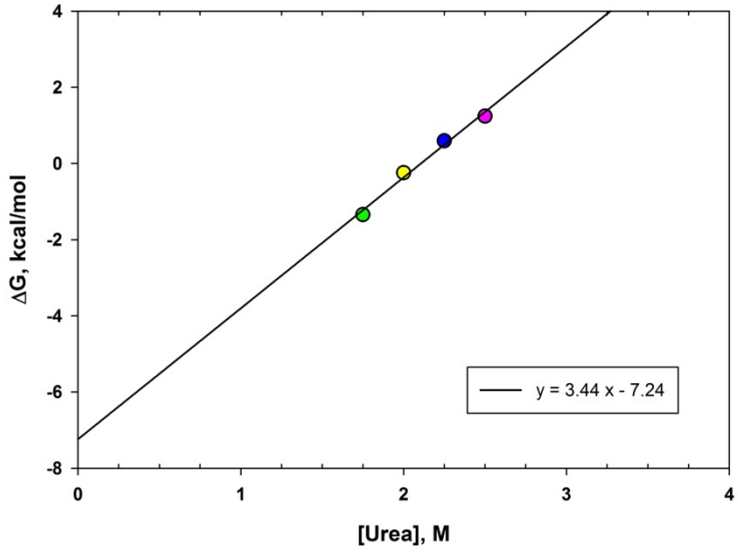


Figure 1.7

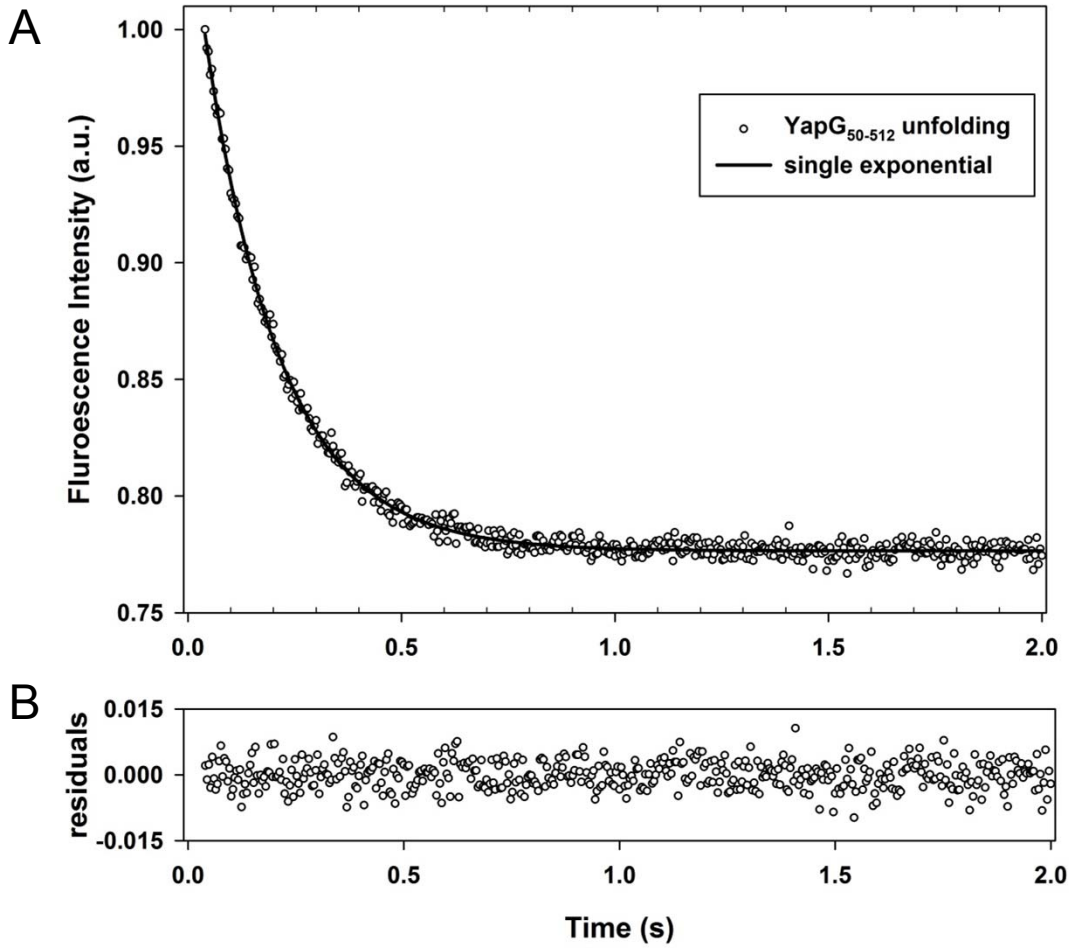

A
B
Fluroescence Intensity (a.u.)
1.00
0.95
0.90
0.85
0.80
0.75
YapG50-512 unfolding
single exponential
0.0
0.5
1.0
1.5
2.0
residuals
0.015
0.000
-0.015
Time (s)

Figure 1.8

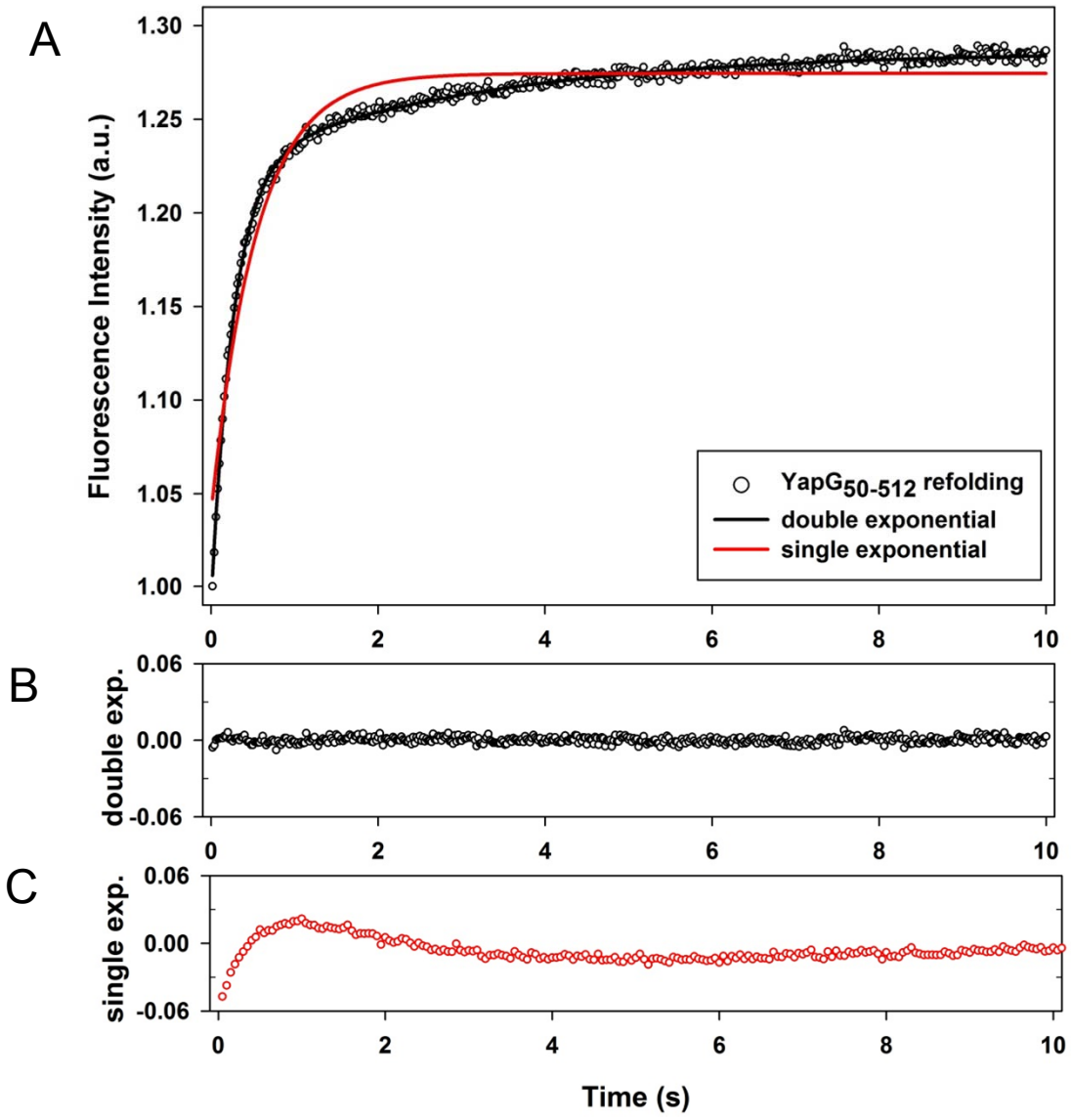
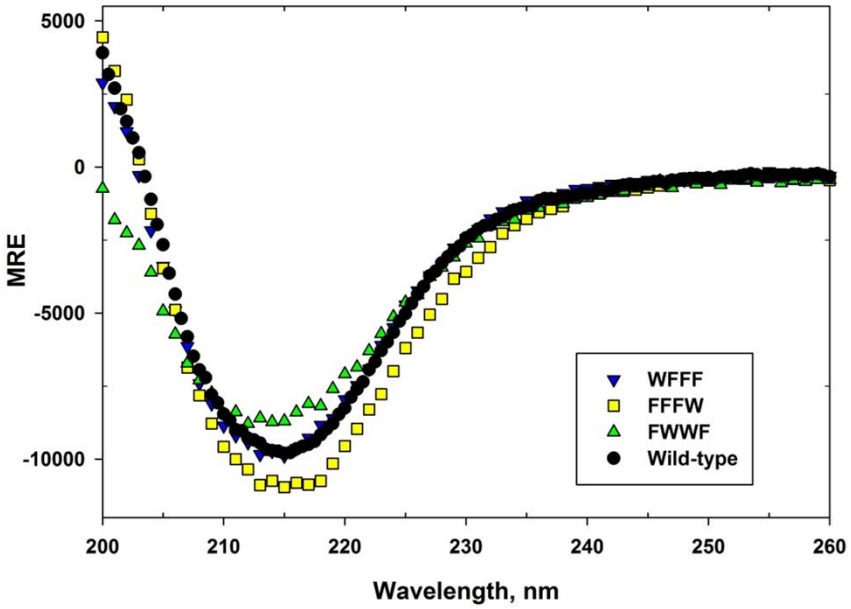

A
B
C
Fluorescence Intensity (a.u.)
1.30
1.25
1.20
1.15
1.10
1.05
1.00
YapG50-512 refolding
double exponential
single exponential
0
2
4
6
8
10
double exp.
0.06
0.00
-0.06
single exp.
0.06
0.00
-0.06
Time (s)

Figure 1.9

A


5000
0
-5000
-10000
MRE
200
210
220
230
240
250
260
Wavelength, nm
WFFF
FFFW
FWWF
Wild-type

B

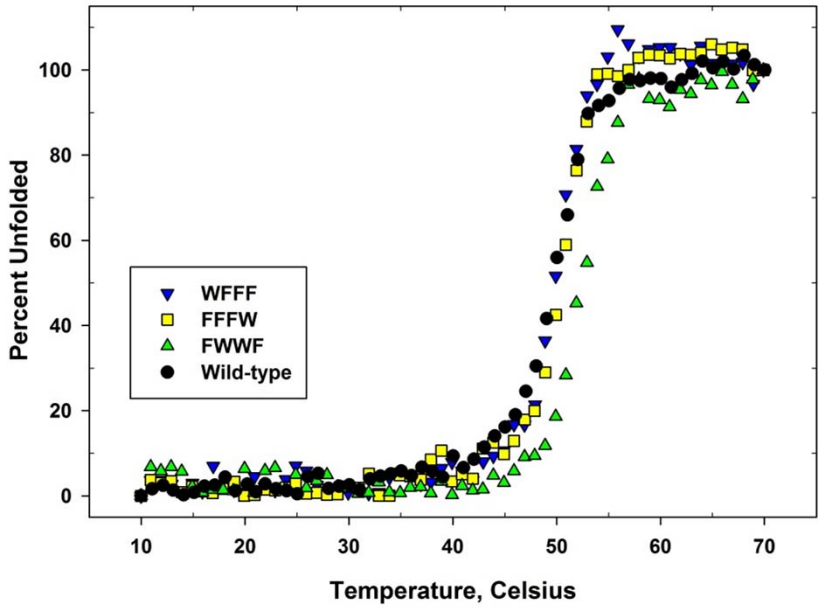
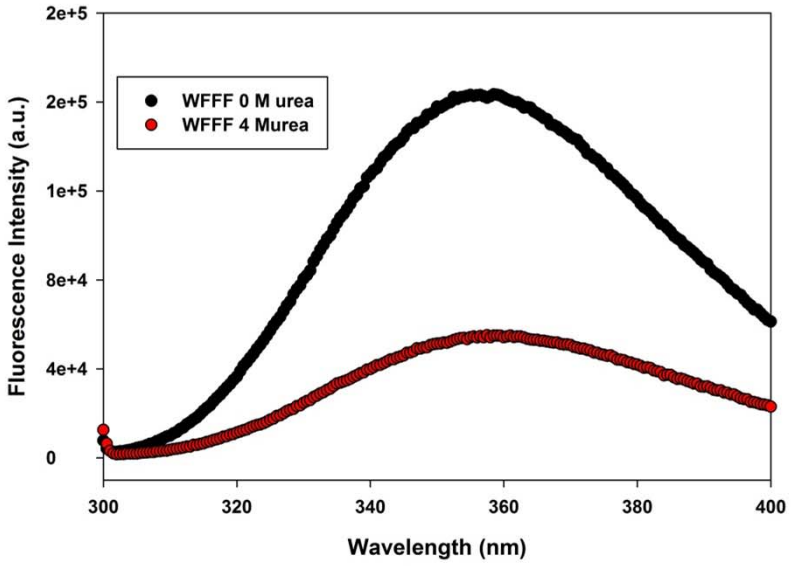

100
80
60
40
20
0
Percent Unfolded
10
20
30
40
50
60
70
Temperature, Celsius
WFFF
FFFW
FWWF
Wild-type

Figure 1.10

A



B

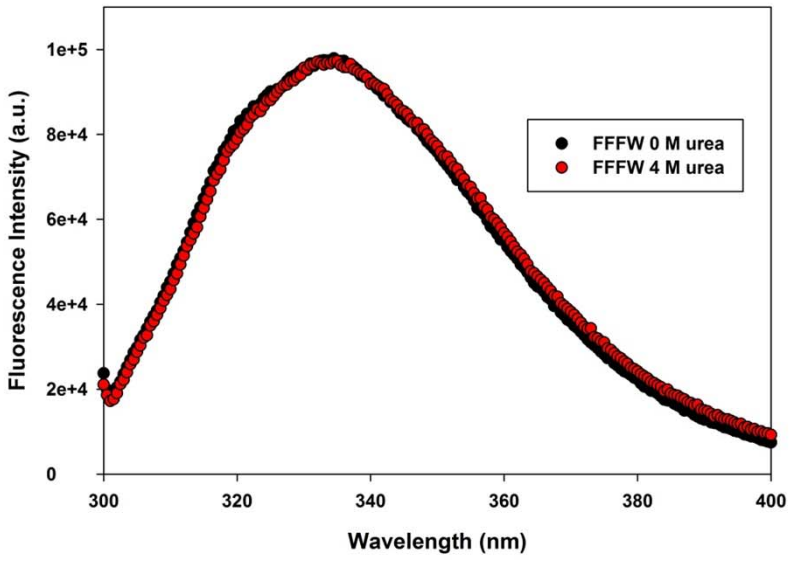


Figure 1.11

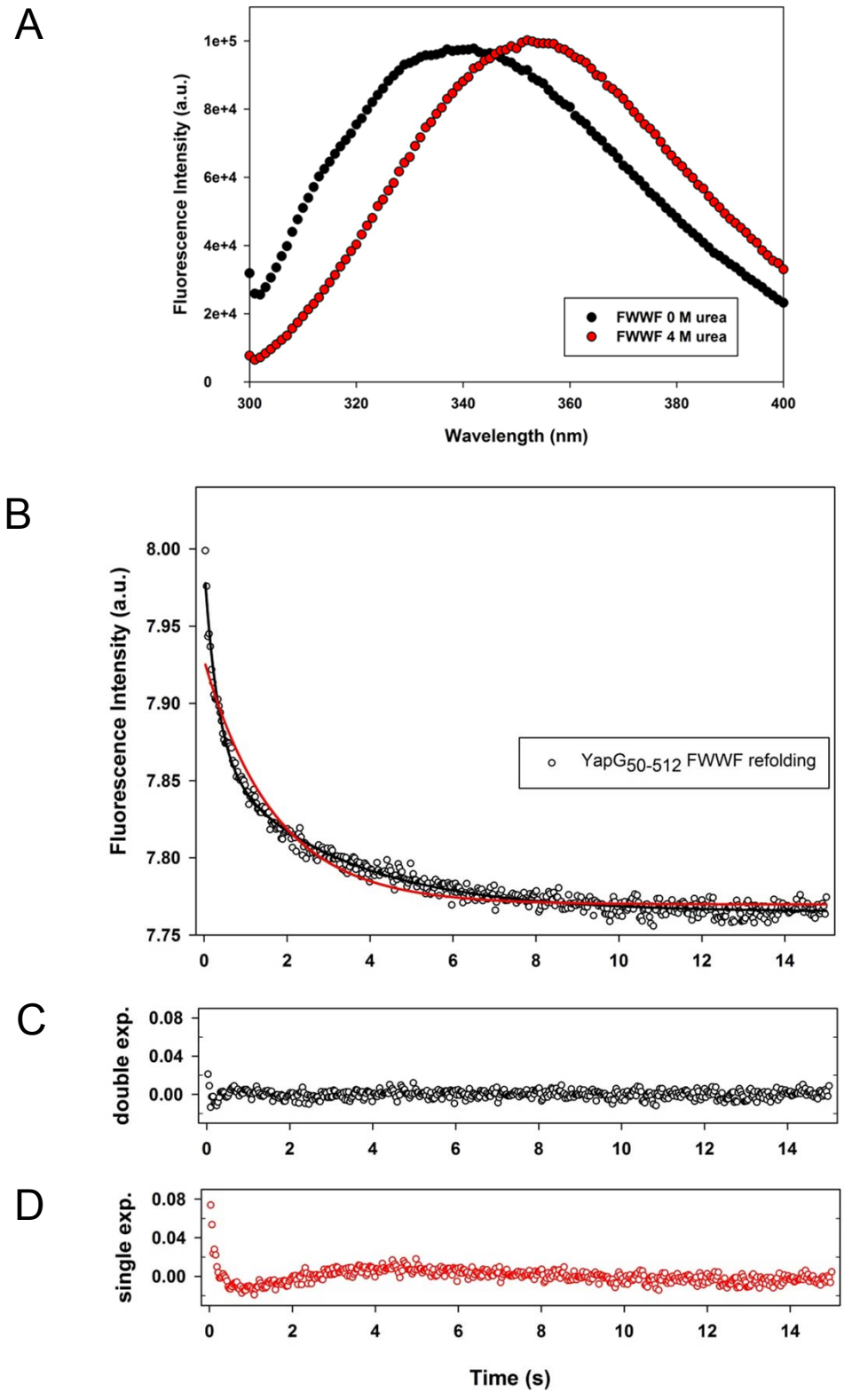

A
Fluorescence Intensity (a.u.)
1e+5
8e+4
6e+4
4e+4
2e+4
0
300
320
340
360
380
400
Wavelength (nm)
FWWF 0 M urea
FWWF 4 M urea
B
Fluorescence Intensity (a.u.)
8.00
7.95
7.90
7.85
7.80
7.75
0
2
4
6
8
10
12
14
YapG50-512 FWWF refolding
C
double exp.
0.08
0.04
0.00
D
single exp.
0.08
0.04
0.00
Time (s)

Figure 1.12

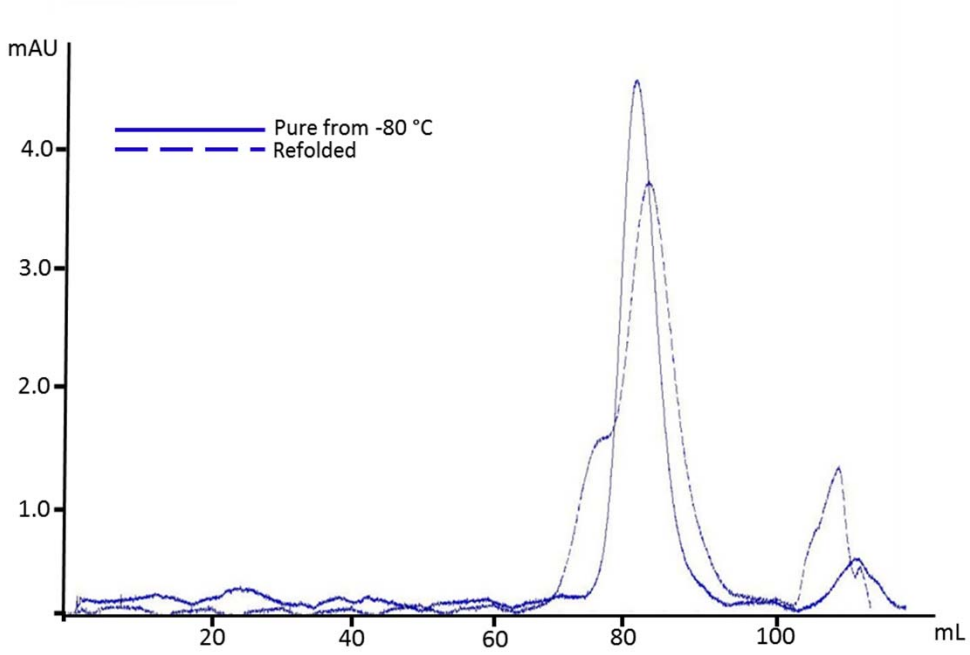


Figure 1.13

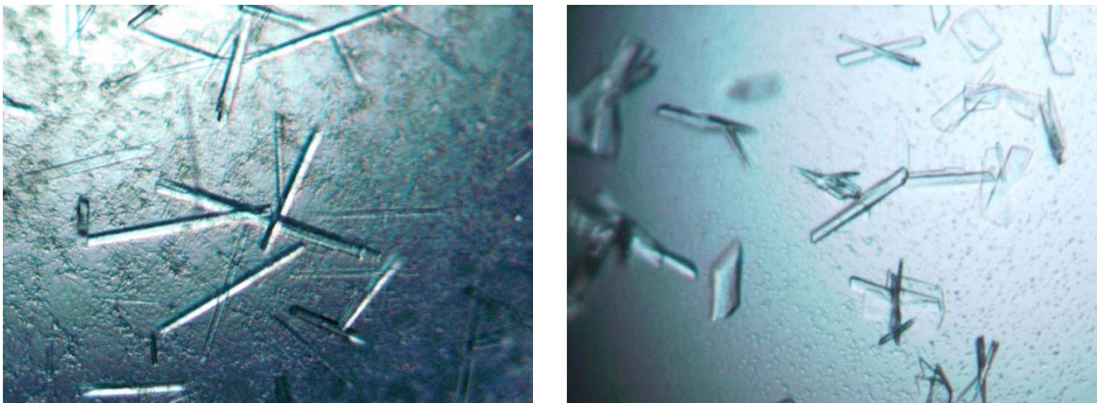
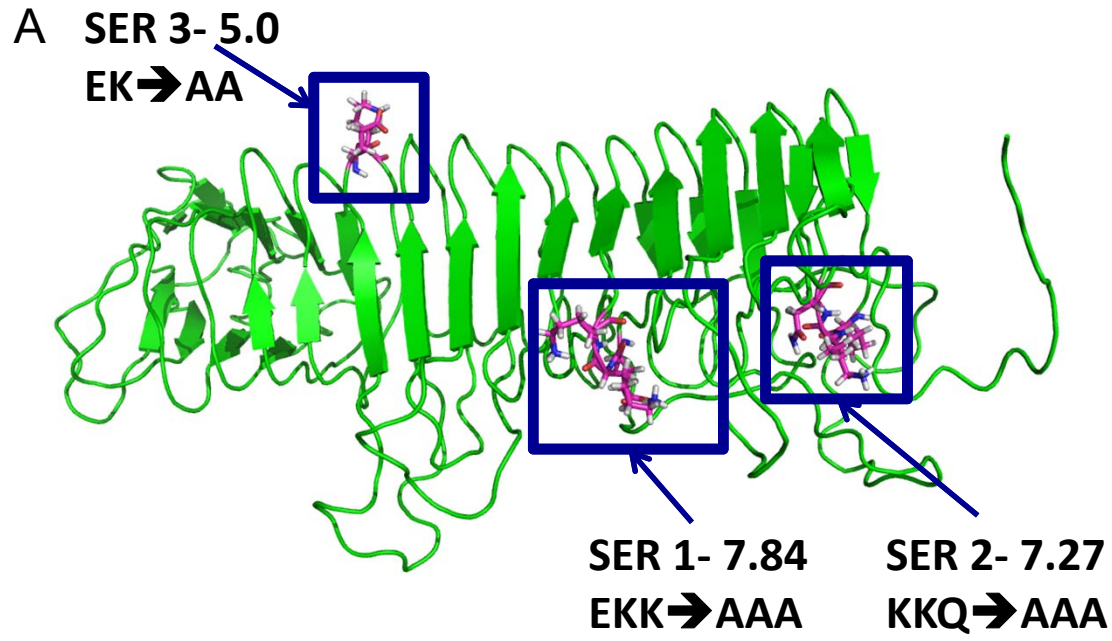
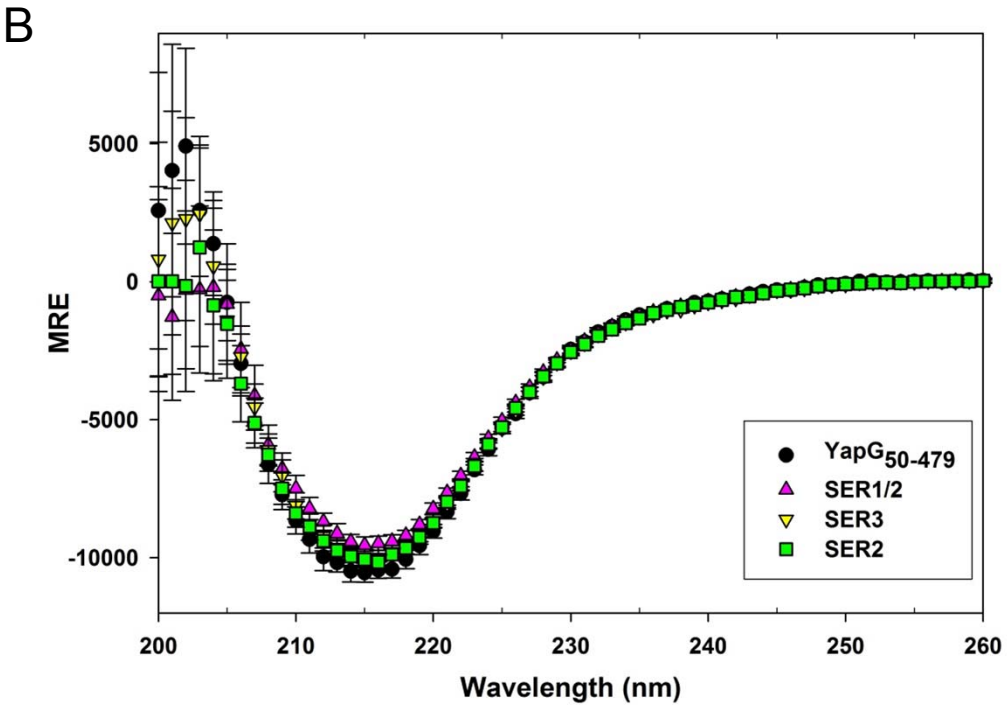


Figure 1.14


A
SER 3- 5.0
EK➔AA
SER 1- 7.84
EKK➔AAA
SER 2- 7.27
KKQ➔AAA


B
5000
0
-5000
-10000
MRE
200
210
220
230
240
250
260
Wavelength (nm)
YapG50-479
SER1/2
SER3
SER2

1.6 REFERENCES

1. Henderson, I. R., Navarro-Garcia, F., and Nataro, J. P. (1998) The great escape: structure and function of the autotransporter proteins, *Trends Microbiol 6*, 370-378.

2. Henderson, I. R., Navarro-Garcia, F., Desvaux, M., Fernandez, R. C., and Ala'Aldeen, D. (2004) Type V protein secretion pathway: the autotransporter story, *Microbiol Mol Biol Rev 68*, 692-744.

3. Jong, W. S., ten Hagen-Jongman, C. M., den Blaauwen, T., Slotboom, D. J., Tame, J. R., Wickstrom, D., de Gier, J. W., Otto, B. R., and Luirink, J. (2007) Limited tolerance towards folded elements during secretion of the autotransporter Hbp, *Mol Microbiol 63*, 1524-1536.

4. Leyton, D. L., Sevastsyanovich, Y. R., Browning, D. F., Rossiter, A. E., Wells, T. J., Fitzpatrick, R. E., Overduin, M., Cunningham, A. F., and Henderson, I. R. (2011) Size and conformation limits to secretion of disulfide-bonded loops in autotransporter proteins, *J Biol Chem 286*, 42283-42291.

5. Sauri, A., Soprova, Z., Wickstrom, D., de Gier, J. W., Van der Schors, R. C., Smit, A. B., Jong, W. S., and Luirink, J. (2009) The Bam (Omp85) complex is involved in secretion of the autotransporter haemoglobin protease, *Microbiology 155*, 3982-3991.

6. Henderson, I. R., and Nataro, J. P. (2001) Virulence functions of autotransporter proteins, *Infect Immun 69*, 1231-1243.

7. Junker, M., Schuster, C. C., McDonnell, A. V., Sorg, K. A., Finn, M. C., Berger, B., and Clark, P. L. (2006) Pertactin beta-helix folding mechanism suggests common themes for the secretion and folding of autotransporter proteins, *Proc Natl Acad Sci U S A 103*, 4918-4923.

8. Yoder, M. D., and Jurnak, F. (1995) Protein motifs. 3. The parallel beta helix and other coiled folds, *Faseb J 9*, 335-342.

9. Otto, B. R., Sijbrandi, R., Luirink, J., Oudega, B., Heddle, J. G., Mizutani, K., Park, S. Y., and Tame, J. R. (2005) Crystal structure of hemoglobin protease, a heme binding autotransporter protein from pathogenic Escherichia coli, *J Biol Chem 280*, 17339-17345.

10. Emsley, P., Charles, I. G., Fairweather, N. F., and Isaacs, N. W. (1996) Structure of *Bordetella pertussis* virulence factor P.69 pertactin, *Nature 381*, 90-92.

11. Khan, S., Mian, H. S., Sandercock, L. E., Chirgadze, N. Y., and Pai, E. F. (2011) Crystal structure of the passenger domain of the *Escherichia coli* autotransporter EspP, *J Mol Biol 413*, 985-1000.

12. Johnson, T. A., Qiu, J., Plaut, A. G., and Holyoak, T. (2009) Active-site gating regulates substrate selectivity in a chymotrypsin-like serine protease the structure of *Haemophilus influenzae* immunoglobulin A1 protease, *J Mol Biol 389*, 559-574.

13. Gangwer, K. A., Mushrush, D. J., Stauff, D. L., Spiller, B., McClain, M. S., Cover, T. L., and Lacy, D. B. (2007) Crystal structure of the *Helicobacter pylori* vacuolating toxin p55 domain, *Proc Natl Acad Sci U S A 104*, 16293-16298.

14. Renn, J. P., and Clark, P. L. (2008) A conserved stable core structure in the passenger domain beta-helix of autotransporter virulence proteins, *Biopolymers 89*, 420-427.

15. Junker, M., and Clark, P. L. (2010) Slow formation of aggregation-resistant beta-sheet folding intermediates, *Proteins 78*, 812-824.

16. Spatara, M. L., Roberts, C. J., and Robinson, A. S. (2009) Kinetic folding studies of the P22 tailspike beta-helix domain reveal multiple unfolded states, *Biophys Chem 141*, 214-221.

17. Fuchs, A., Seiderer, C., and Seckler, R. (1991) In vitro folding pathway of phage P22 tailspike protein, *Biochemistry 30*, 6598-6604.

18. Kamen, D. E., Griko, Y., and Woody, R. W. (2000) The stability, structural organization, and denaturation of pectate lyase C, a parallel beta-helix protein, *Biochemistry 39*, 15932-15943.

19. Lenz, J. D., Lawrenz, M. B., Cotter, D. G., Lane, M. C., Gonzalez, R. J., Palacios, M., and Miller, V. L. (2011) Expression during host infection and localization of *Yersinia pestis* autotransporter proteins (Yaps), *J Bacteriol*.

20. Lawrenz, M. B., Lenz, J. D., and Miller, V. L. (2009) A novel autotransporter adhesin is required for efficient colonization during bubonic plague, *Infect Immun 77*, 317-326.

21. McCarter, J. D., Stephens, D., Shoemaker, K., Rosenberg, S., Kirsch, J. F., and Georgiou, G. (2004) Substrate specificity of the *Escherichia coli* outer membrane protease OmpT, *J Bacteriol 186*, 5919-5925.

22. Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol 340*, 783-795.

23. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Eng 10*, 1-6.

24. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) The Pfam protein families database, *Nucleic Acids Res 38*, D211-222.

25. Williamson, M. P. (1994) The structure and function of proline-rich regions in proteins, *Biochem J 297 (Pt 2)*, 249-260.

26. Bryson, K., McGuffin, L. J., Marsden, R. L., Ward, J. J., Sodhi, J. S., and Jones, D. T. (2005) Protein structure prediction servers at University College London, *Nucleic Acids Res 33*, W36-38.

27. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol 292*, 195-202.

28. Cole, C., Barber, J. D., and Barton, G. J. (2008) The Jpred 3 secondary structure prediction server, *Nucleic Acids Res 36*, W197-201.

29. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. (1998) JPred: a consensus secondary structure prediction server, *Bioinformatics 14*, 892-893.

30. Rost, B., Yachdav, G., and Liu, J. (2004) The PredictProtein server, *Nucleic Acids Res 32*, W321-326.

31. Bradley, P., Cowen, L., Menke, M., King, J., and Berger, B. (2001) BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens, *Proc Natl Acad Sci U S A 98*, 14819-14824.

32. Kamen, D. E., and Woody, R. W. (2002) Folding kinetics of the protein pectate lyase C reveal fast-forming intermediates and slow proline isomerization, *Biochemistry 41*, 4713-4723.

33. Miller, S., Schuler, B., and Seckler, R. (1998) A reversibly unfolding fragment of P22 tailspike protein with native structure: the isolated beta-helix domain, *Biochemistry 37*, 9160-9168.

34. Greenfield, N. J. (2006) Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism, *Nat Protoc 1*, 2733-2741.

35. Greenfield, N. J. (2006) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions, *Nat Protoc 1*, 2527-2535.

36. Goldschmidt, L., Cooper, D. R., Derewenda, Z. S., and Eisenberg, D. (2007) Toward rational protein crystallization: A Web server for the design of crystallizable protein variants, *Protein Sci 16*, 1569-1576.

37. Pakula, A. A., and Sauer, R. T. (1989) Genetic analysis of protein stability and function, *Annu Rev Genet 23*, 289-310.

38. Plaxco, K. W., Simons, K. T., and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins, *J Mol Biol 277*, 985-994.

39. Schmid, F. X. (1995) Protein folding. Prolyl isomerases join the fold, *Curr Biol 5*, 993-994.

40. Schmid, F. X. (1986) Fast-folding and slow-folding forms of unfolded proteins, *Methods Enzymol 131*, 70-82.

41. Brandts, J. F., Halvorson, H. R., and Brennan, M. (1975) Consideration of the Possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues, *Biochemistry 14*, 4953-4963.

42. Soding, J., Biegert, A., and Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction, *Nucleic Acids Res 33*, W244-248.

43. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2007) Comparative protein structure modeling using MODELLER, *Curr Protoc Protein Sci Chapter 2*, Unit 2 9.

44. Stols, L., Gu, M., Dieckman, L., Raffen, R., Collart, F. R., and Donnelly, M. I. (2002) A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site, *Protein Expr Purif 25*, 8-15.

45. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Int J Neural Syst 8*, 581-599.

CHAPTER 2

Crystal Structure of the Plant Epigenetic Protein Arginine Methyltransferase 10

2.1 INTRODUCTION

Protein arginine methyltransferases (PRMTs) catalyze the transfer of methyl groups from S-adenosylmethionine (SAM) to arginine residues of target proteins, and release S-adenosylhomocysteine (SAH) as a product (*1*). The post-translational methylation of arginines is observed widely in eukaryotes and plays essential roles in many biological processes, such as signal transduction, chromatin remodeling, RNA processing, gene transcription, DNA repair and cellular transport (*1-8*). PRMTs are generally classified as type I or type II (*1*). Both types catalyze the production of ω-N^G-monomethylarginine, but they generate distinct dimethyl arginine derivatives. Type I enzymes (*e.g.,* PRMT1, 3, 4, 6, 8) specifically produce asymmetric ω-N^G,N'^G-dimethylarginine, while type II enzymes (*e.g.,* PRMT5, 7 and FBXO11) only produce symmetric dimethylarginine (*9*). The dysfunctions of mammalian PRMTs have been correlated with the development of cancer as well as autoimmune, cardiovascular, pulmonary and neuro-developmental diseases (*10-16*).

While PRMTs have a relatively conserved catalytic core, the portions of each

Reprinted from Journal of Molecular Biology, 414(1), Yuan Cheng, Monica Frazier, Falong Lu, Xiaofeng Cao, and Matthew R. Redinbo, Crystal Structure of the Plant Epigenetic Protein Arginine Methyltransferase 10, 106-22, 2011, with permission from Elsevier.

Monica Frazier contributed Section 2.3.6, a portion of Section 2.4, Figure 2.8, Figure 2.9, and Supplemental Figure 2.3.

enzyme N-terminal to the catalytic core (the “N-terminal additions”) are divergent in sequence and have been demonstrated to be important for the substrate specificity. For example, the zinc-finger domain within the N-terminal addition of PRMT3 is essential for its recognition of RNA-associated targets (*17*). Previous structural studies have shown that the PRMT catalytic core is composed of three domains: an N-terminal SAM binding domain, a central arm domain, and a C-terminal β-barrel domain (*18*). The main substrate binding site is located in a cleft formed between the SAM binding domain and the β-barrel domain (*19, 20*). Dimerization is a conserved feature in PRMTs and has been established to be essential for the methyltransferase activity of PRMTs (*19, 20*) by facilitating SAM binding (*20*).

PRMT methyltransferase activity is regulated by several characteristics of the target protein. For example, the local sequence of the methylation site is an important determinant of arginine methylation (*21, 22*). PRMT-catalyzed reactions typically occur within glycine- and arginine-rich motifs, such as “RG”, “RGG” and “RXR” (*23*), although exceptions have been noted (*22*). The activity of PRMTs can also be affected by the sequences distal to the methylation site (*24*) and by protein binding partners (*25, 26*). Circumstantial evidence has suggested that PRMTs often form complexes with other proteins *in vivo,* and that these proteins impact subcellular location and substrate recognition (*27, 28*).

AtPRMT10 is a plant-specific type I PRMT that plays an essential role in the regulation of flowering time in *Arabidopsis* (*29*). Genetic disruption of *AtPRMT10* causes delayed flowering due to up-regulated transcription of a major flowering repressor, *FLOWERING LOCUS C* (*FLC*) (*29*). Biochemical studies showed that AtPRMT10 can specifically methylate arginine-3 of both histone H4 and histone H2A *in vitro,* and preferentially produces asymmetrical dimethylarginines. Besides AtPRMT10, eight other AtPRMTs have been identified in the *Arabidopsis* genome, including AtPRMT1a, AtPRMT1b, AtPRMT3, AtPRMT4a, AtPRMT4b, AtPRMT5, AtPRMT6 and AtPRMT7. These

AtPRMT paralogs likely have diverse properties in cellular location, substrate specificity and protein-protein interaction (*1, 7*).

Here we report the crystal structure of AtPRMT10 in complex with a product of its enzymatic reaction, SAH. This structure provides insights into how AtPRMT10 interacts with peptides, and reveals structural features that may confer unique substrate specificity to AtPRMT10, including the role of the AtPRMT10 N-terminal addition in the enzyme function. Our studies also show that AtPRMT10 exists predominantly in a dimeric state in solution, and disruption of dimerization causes loss of activity. We further examine the impact AtPRMT10 dimerization has on enzyme motion using molecular dynamics (MD) simulations. Our results highlight distinct differences between AtPRMT10 and other structurally-characterized PRMTs, but also indicate that motions are a conserved element of PRMT function.

2.2 RESULTS

2.2.1 Crystal Structure of the AtPRMT10-SAH Complex

The structure of AtPRMT10 (residues 11-383) in complex with SAH was determined by molecular replacement and refined to 2.6 Å resolution (Table 2.1). The crystal specimen employed to solve the structure contained nearly 50% pseudomerohedral twinning as indicated by the L-test and Britton plot carried out by the program PHENIX (*30*). Notably, the β angle (89.98^{o}) of the unit cell was very close to 90^{o}. Consequently, the diffraction data could also be reduced into the orthorhombic space group P222 and its derivatives. Serious violations of systematic absences were observed, however, in space groups $P2_12_12_1$, $P2_12_12$, $P222_1$. Consequently, molecular replacement was performed in the space group P222. As predicted by the Matthews coefficient, two monomers were identified in each asymmetric unit. All solutions in space group P222, however, were finally rejected owing to

the presence of significant main-chain clashes during crystal packing. Taken together, these results indicated that $P2_1$ was the correct space group and pesudomerohedral twinning was present. The data were detwinned using the twining refinement function of the program PHENIX (*30*). No violations in systematic absences for $P2_1$ and no clashes between protein monomers with four AtPRMT10-SAH complexes per asymmetric unit were observed; the structure was determined and refined with good geometry and statistics in this space group (Table 2.1). The N-terminal twenty residues of the AtPRMT10 construct employed (residues 11-30) lack electron density and were not placed in the final refined model.

AtPRMT10 exhibits three sequentially folded domains: an N-terminal SAM binding domain (residue 31-174), a central arm domain (residues 187-236), and a C-terminal β-barrel domain (residues 175-186 and residues 237-383) (Figure 2.1a, b). The SAM binding domain is composed of two N-terminal helices (αX & αY, residues 31-50) followed by a classical Rossman fold (residues 51-174) consisting of five α helices (αZ, αZ', αA, αB, αD) and five β strands (β1 to β5). The consensus Rossman fold has been observed in other known SAM-dependent methyltransferases (*31, 32*), while the two N-terminal helices (αX & αY) are unique to PRMTs (*19*). The β-barrel domain, forming close contacts with the SAM-binding domain at one end of its barrel, harbors ten β-strands (β6 to β15) and two short α-helices (αH and αI). The arm domain, exhibiting a helix-turn-helix fold, is inserted in between β6 and β7 of the β-barrel domain and protrudes from the main body of the protein. Sequence analysis reveals four PRMT signature motifs in AtPRMT10 (Figure 2.2). Motif I (YFxxY) and Motif II (DVGxGxG) are directly involved in the binding of cofactor SAM. Motif III (SExMGxxLxxExM), harbors two critical catalytic residues E143 and E152. Mutation of either of these two residues completely disrupted the methyltransferase activity of AtPRMT10 (data not shown). Motif IV (or the THW motif) is the most highly conserved sequence among PRMTs and is directly involved in the formation of the active site. As

expected, disruption of motif IV is accompanied with complete loss of the methyltransferase activity of AtPRMT10 (data not shown).

The structure of AtPRMT10 exhibits a similar overall fold relative to other PRMTs of known structure, exhibiting, for example, a 1.8 Å root-mean-square deviation over 245 Cα positions with PRMT1 (residues 41-354). However, a strikingly unique feature of AtPRMT10 is its dimerization arm, consisting of two straight anti-parallel α-helices, which is significantly longer (41 Å) than that of other PRMTs (*e.g.,* PRMT1, 22 Å; PRMT3, 22 Å; CARM1, 34 Å) (Figure 2.1c). AtPRMT10 also differs from other PRMTs in two loop regions of the β-barrel domain (Figure 2.1c). Sequence alignment indicates that these loops are relatively conserved among AtPRMT10 orthologs (Supplemental Figure 2.1), but highly divergent among PRMT paralogs (Figure 2.2). Loop I is located adjacent to a conserved substrate binding site of PRMTs (see below). Acidic residues in Loop II have been shown to be important for the interaction of PRMT1 with its substrates (*33*).

2.2.2 AtPRMT10 Active Site

In the AtPRMT10-SAH complex, SAH binds within a deep pocket formed by the three N-terminal α-helices (αX, αY and αZ) and the carboxyl ends of the parallel β-strands (β1 to β5) (Figure 2.1d). Most of the residues involved in SAH binding are highly conserved among type I PRMTs (Figure 2.2), indicating that members of the type I PRMT family likely share similar mechanisms in cofactor binding and catalysis. Hydrogen bonding plays a major role in the interaction of AtPRMT10 with SAH, with six such interactions formed between AtPRMT10 and the three moieties of SAH (adenine, ribose and homocysteine). R54 of the helix αZ forms bifurcated hydrogen bonds with the terminal carboxylate group of the homocysteine moiety. For the ribose moiety, hydrogen bonds are observed between the two main-chain hydroxyl groups and the side chains of E100 of strand β2 and Q45 of helix αY. The adenine group is recognized by the E129 from the loop between β2 and β4. In

addition to hydrogen bonding, the main-chain of the glycine rich loop (G78 and G80) and the side-chains of seven other residues (A101, V128, F36, M154, S157, Y35 and Y39) form van der Waals contacts with SAH. Given the small difference between the chemical structure of SAM and SAH, it is expected that SAM binds to the active site in a manner similar to that observed here for SAH.

2.2.3 AtPRMT10 Dimer

AtPRMT10 forms a ring-like homodimer through the interaction between the dimerization arm (αE-loop-αG) of one monomer and the outer surface (αY, αZ, αA & αD) of the SAM binding domain of the other monomer (Figures 2.3a, b). Both active sites are located at the periphery of the central cavity formed upon dimerization of AtPRMT10. As observed for other PRMTs, hydrophobic interactions are a major force during the formation of the AtPRMT10 dimer. A network of three hydrogen bonds is also observed at the PRMT dimer interface, with the side-chains of Q90 and N115 forming hydrogen bonds with the main-chains of G215 and D217 respectively (Figure 2.3c). The hydrogen bonds between N115 and D217 are highly conserved among PRMTs (Figure 2.2). Another conserved residue at the dimer interface is G215, whose small side-chain is apparently favorable for the formation of the sharp turn at the tip of the dimerization arm. Overall, the residues on the surface of the SAM binding domain that produce the AtPRMT10 dimer interface are highly conserved when compared to other PRMTs. In contrast, however, the residues that form the dimerization arm of AtPRMT10 exhibit little or no conservation with homologous enzymes (Figure 2.2 and Supplemental Figure 2.2).

Notably, due to PRMT's longer dimerization arm, its central cavity is significantly larger than those of other PRMTs with known structure (Figure 2.4). AtPRMT10 creates a cavity 15 Å high by 13 Å wide (15 x 13 Å), while those of PRMT1, PRMT3 and CARM1 exhibit cavities that are 8x12, 8x13 and 8x11 Å, respectively (Figure 2.4). The longer

"vertical" distance as depicted in Figure 2.4 is generated by the longer AtPRMT10 dimerization arm. Consistent with the dimer observed in the crystal structure, our results from dynamic light scattering and gel filtration experiments confirmed that AtPRMT10 exists predominately as dimer in solution, and that the oligomeric state of the enzyme is independent of SAH binding (Table 2.2).

To test the importance of the dimer interface observed in the crystal structure during AtPRMT10 function we designed an arm mutant, Δ203-225, in which the part of the dimerization arm that forms the dimer interface was replaced with a stretch of glycine and serine residues (GGSGGS). AtPRMT10 Δ203-225 was stably over-expressed in *E.coli*, suggesting that it was well folded. The oligomeric state of AtPRMT10 Δ203-225 was examined using dynamic light scattering and gel filtration experiments (Table 2.2). Our results show that mutation of the dimerization arm disrupted dimer formation. The impact of dimerization on the methyltransferase activity of AtPRMT10 was examined by measuring the activity of the arm mutant Δ203-225. The arm mutant displayed no observable activity toward H2A and H4 (Figure 2.5a, b), indicating that dimerization is essential for the methyltransferase activity of AtPRMT10.

2.2.4 AtPRMT10 Surface Electrostatics

Surface charge distribution appears to impact the function of PRMTs. For example, published data have suggested that surface charges are crucial for the interaction of PRMT with substrates and other proteins (*19, 20*). Figure 2.6 illustrates the surface charge distribution of AtPRMT10. As seen in other PRMTs, the surface of AtPRMT10 contains numerous acidic patches, especially around the active site. However, there are notable differences in the surface charge distribution of AtPRMT10 compared to other PRMTs of known structure (Figure 2.6). In particular, the unusually long dimerization arm of AtPRMT10

contains ten acidic residues (E190, D195, D197, D202, D208, E209, D217, E227, E228, E230) (Figure 2.6) that generate a relatively large acidic surface along this domain relative to other PRMTs. A second difference is observed at one end of the β-barrel domain, where AtPRMT10 has a large acidic patch formed by residues E281, E336, E337, D339, E367 and E374 (Figure 2.6). Other PRMTs contain fewer acidic residues in this region (Figure 2.2). Acidic amino acid residues in this location have been shown to be important for the substrate interaction of PRMT1 (*33*).

Structural studies of PRMT1 have indicated the location of the substrate binding groove of this enzyme (*20*). Based on the location of acidic patches and the shape of the AtPRMT10 surface in light of other PRMTs of known structure, we have identified four putative substrate binding grooves on the surface of AtPRMT10 (Figure 2.6). Binding grooves I and II are located in the cleft formed between the SAM binding domain and the β-barrel domain and are directly connected to the active site. Binding grooves III and IV lie on the surface of the β-barrel domain. Substrates can also enter the active site through binding groove III. A high degree of conservation is maintained in the residues that form binding grooves I and II (Figure 2.6 and Supplemental Figure 2.2), suggesting the conserved role for these two binding grooves during substrate interaction. In contrast, little conservation is observed for the residues that form binding grooves III and IV (Figure 2.6 and Supplemental Figure 2.2). It is possible that the unique compositions of binding grooves III and IV may confer unique substrate specificities upon AtPRMT10 compared to other PRMTs.

2.2.5 Increased Active Site Accessibility in AtPRMT10

While the PRMT family shares a three-domain architecture and a dimeric oligomerization state, the relative orientation of the two monomers in a functional dimer significantly varies between different PRMTs due to the diversity in dimerization arm length

and composition. Consequently, the dimeric forms of different PRMTs do not superimpose well. When we align different PRMTs based on one of their two monomers (the “bottom monomers” in Figure 2.7a, left panel), the other monomers (the “top monomers”) are translated to distinct locations. In Figure 7a (left panel), the top monomers of PRMT1 (cyan) and PRMT3 (yellow) are located directly above their bottom monomers, while the top monomers of AtPRMT10 (magenta) and CARM1 (blue) are positioned away from the vertical by 30^{o} and 20^{o}, respectively. The top monomers of AtPRMT10 and CARM1 are also observed to be translated leftward 21 Å and 13 Å to the left, respectively, relative to the position of PRMT1 and PRMT3 (Figure 2.7a, middle panel). Finally, the angles formed by the two monomers of a PRMT dimer vary significantly among enzyme paralogs, ranging from 30^{o} in PRMT3 to 52^{o} in AtPRMT10 (shown schematically in Figure 2.7a, right panel).

The differences in the relative orientation of the two monomers in PRMT dimers, together with the differences in the size of the central enzyme cavities, result in significant variations in active site accessibility across the enzymes of known structure. To provide a quantitative measure of active sites accessibility for different PRMTs, we determined an accessibility angle for AtPRMT10, CARM1, PRMT1 and PRMT3. With the bottom monomers in the same orientation, a vertex was placed in the center of dimer cavity, and from that point the largest angle allowed by the molecular surface of the dimer in two dimensions in this view was traced for each structure (Figures 2.7b-7e). For AtPRMT10, the accessibility angle was ~120° (Figure 2.7b). However, for PRMT1 and PRMT3 and CARM1, the accessibility angles were ~50°, ~45°, and ~20°, respectively (Figures 2.7c-7e). Thus, the unique size and orientation of the AtPRMT10 central cavity creates a significantly larger accessibility to this enzyme’s active site relative to the PRMTs of currently known structure.

2.2.6 AtPRMT10 Motion

Because dimerization has been shown to be essential for the methyltransferase activity of PRMTs, we examined the impact dimerization has on the motion of AtPRMT10 using 30 ns molecular dynamics (MD) simulations. Monomeric and dimeric forms of AtPRMT10 were examined. The total energy of each system, calculated as the sum of the kinetic and potential energy at each time point, was relatively constant after the first 5 ns, particularly over the last 10 ns (Figure 2.8). Therefore, the averages of the MD trajectories in the last 10 ns were used for the following analysis. The effect of dimerization on the degree of motion of AtPRMT10 was determined by computing the atomic position fluctuations (APFs) of Cα atoms of the monomer and dimer form. Overall, AtPRMT10 exhibits similar APFs in monomeric and dimeric states; however, in the dimeric form, two regions (αY-loop-αZ, residues 40-68; the dimerization arm, residues 187-235) displayed significantly lower APFs than when in the monomeric form (Figures 2.9a, b). The reduced fluctuations within these two regions likely result from their direct involvement in the formation of the dimer interface (Figure 2.3C). Notably, the region αY-loop-αZ (residues 40-68) is directly involved in the binding of SAH and in the formation of substrate binding groove I. Therefore, stabilization of this region by dimerization likely improves the binding of SAH and substrate proteins.

We computed normalized covariance matrices to classify the motions of all residue pairs in the protein (Figures 2.9c, d). Normalized covariance matrices generate the residue-residue correlation coefficients (C_{ij}s), which inform the relative motion between a residual pair. Based on the value of C_{ij}s, the motions of all residue pairs can be classified into three groups: correlated motion (two residues moving toward the same direction) as indicated by C_{ij} approaching 1, anti-correlated motion (two residues moving toward the opposite direction) as indicated by C_{ij} approaching -1, and uncorrelated motion (two residues moving with the lack of a dynamic relationship) with C_{ij} values near zero. The SAM binding domain

of dimeric AtPRMT10 exhibits considerably greater residue-residue correlations relative to that of monomeric AtPRMT10 (Figure 2.9c). Increased residue-residue correlations are also observed in several discrete regions of the β-barrel domain.

To better understand the biological significance of residue-residue correlations, single-linkage clustering analysis was then conducted to identify groups of residues that move together. Clustering of dimeric AtPRMT10 at a correlation coefficient above 0.7 resulted in five clusters, while clustering of monomeric AtPRMT10 under the same criterion only resulted in three clusters (Figures 2.9e, f). One notable difference between monomeric AtPRMT10 and dimeric AtPRMT10 lies in the SAM binding domain. Most of this region, except the two N-terminal helices (αX and αY) and two loop regions (L1 and L2), are clustered in dimeric AtPRMT10 (Figure 2.9g), while only helix B is self-clustered in monomeric AtPRMT10. In addition, one end of the β-barrel domain is clustered in dimeric AtPRMT10, but not in monomer AtPRMT10. These data establish that the SAM binding domain and one end of the β-barrel domain to move as a cohesive unit in dimeric AtPRMT10, but not in monomeric AtPRMT10.

To extend these investigations into other PRMTs, we examined the motion of monomeric and dimeric PRMT3 using the same MD simulation protocol described above. Similar to AtPRMT10, dimerization significantly lowered the APFs in the N-terminal region (αX-αY-αZ, residues 208-245) and the dimerization arm (residues 370-394) (Supplemental Figure 2.3a). In addition, normalized covariance analysis clearly shows that dimerization promotes coherent protein motions in the SAM binding domain and several discrete regions of the β-barrel domain (Supplemental Figure 2.3b, c). Taken together, our results show that dimerization productively impacts the motion of the PRMTs. In particular, the SAM binding domain in both AtPRMT10 and PRMT3 move as a cohesive unit in the enzyme dimer but not the monomer.

2.2.7 PRMT10 N-terminus in Enzyme Function

Finally, we examined the impact of the N-terminal addition (residues 1-30) on the dimeric state and methyltransferase activity of AtPRMT10. We created three N-terminal deletion mutants, including ΔN10 (residues 11-383), ΔN20 (residues 21-383) and ΔN30 (residues 31-383), and compared their biophysical properties and methyltransferase activities to those of full-length AtPRMT10. The oligomeric states of these mutants were investigated using dynamic light scattering (DLS) and gel filtration experiments (Table 2.2). As observed in the wild-type enzyme, all N-terminal deletion mutants form dimers in solution. Moreover, the oligomeric states of these mutants are SAH-independent. Together, these data suggest that the N-terminal addition does not impact AtPRMT10 dimerization.

The methyltransferase activities of wild-type AtPRMT10 and the three N-terminal deletion mutants were measured as described previously (*29*); while these initial studies do not provide kinetic values, they are sufficient to highlight relative differences in enzyme function (Figure 2.5a, b). Purified calf thymus core histones, which are a mixture of histones H2A, H2B, H3 and H4, were chosen as the substrate. Of these four histones, H2A and H4 are known to be methylated by AtPRMT10. Upon analysis of the experiment by SDS-PAGE, the methylation state of H2A and H4 can be quantified individually, due to their difference in molecular weight. Interestingly, ΔN10 had approximately 3-fold greater activities toward H2A relative to the wild-type enzyme. Additional deletions of the N-terminus (ΔN20 and ΔN30) did not enhance the methylation of H2A by AtPRMT10. When H4 was used as the substrate, however, all three N-terminus mutants displayed wild-type level activities (Figure 2.5a, b). These results indicate that first ten residues of AtPRMT10 impact enzyme methyltransferase activity in a protein substrate-dependent manner. In the AtPRMT10 crystal structure, the helix αX (residues 32-40) covers the opening of the SAM binding pocket and stabilizes SAH binding with van der Waals interactions (Figure 2.1). As expected, the deletion of helix αX (Δ40) causes a dramatic drop in the activity for both H2A and H4.

Previous studies of PRMT1 have shown that amino acids distal to the methylation site can affect the methylation of H4 (*24*). Thus, we examined whether the substrate sequence outside of the methylation site also impacts the activity of AtPRMT10. We examined the purified full-length histone H4 as well as H4N1-20, a peptide covering only the N-terminal twenty residues of histone H4. We found that the activity of AtPRMT10 on the full-length H4 substrate was markedly higher than that on the H4N1-20 substrate (Figure 2.5c), in spite of a 10-fold higher concentration of H4N1-20 was present in these assays. Thus, it appeared that the sequence downstream of the N-terminal 20 residues of histone H4 enhanced the methyltransferase activity of AtPRMT10.

The methylation site of AtPRMT10 in both histone H2A and histone H4 is located at the far N-terminus of these proteins. To examine whether a bulky protein fused to the N-terminus of H4 would impact AtPRMT10 activity at arginine-3 on H4, we compared the methylation of histone H4 and N-terminally GST-tagged histone H4 (GST-H4) by AtPRMT10 (Figure 2.5c). Our results show that the presence of a N-terminal GST tag modestly reduced the activity of AtPRMT10 by ~2-fold relative to untagged H4. These data indicate that AtPRMT10 can methylate R3 of H4 even when it is not located at the far N-terminus of this histone protein.

2.3 DISCUSSION

We present the first structure of a plant protein arginine methyltransferase, that of AtPRMT10, and highlight unique features of this enzyme, including a long dimerization arm and a distinctly open conformation in the catalytic dimer. We also establish for the first time that the family of PRMTs exhibit conserved domain motions, particularly within the enzyme region that binds the SAM cofactor that donates the methyl group to arginines on target proteins. Together, these data advance our understanding of features shared by the PRMT

enzymes, which function as both epigenetic and non-epigenetic factors, as well as unique aspects particular family members may employ to impact substrate preference.

In a functional PRMT dimer, the enzyme active sites are located at the periphery of a central cavity (Figure 2.3). This configuration likely impacts access of substrate proteins to the PRMT catalytic site. Indeed, most known methylation sites are located in disordered regions of substrates, and the structural flexibility around the methylation site has been shown to be essential for PRMT function (*22*). Comparing PRMT dimers of known structure demonstrates that PRMT paralogs exhibit a range of accessibility in the active site, which can be roughly summarized as: AtPRMT10>PRMT1>PRMT3>CARM1 (Figure 2.7). These variations result primarily from differences in the relative orientation of the two monomers in a functional PRMT dimer and differences in the dimerization arm length. Previous studies have suggested that the activity and substrate specificity of PRMTs are directly correlated with active site accessibility (*34*). Thus, it is possible that the more accessible AtPRMT10 active site may allow this enzyme to methylate arginine residues that do not serve as substrates for other PRMTs.

We show that AtPRMT10, like other PRMTs, functions only as a dimer (Figure 2.5). MD simulations on both the monomeric and dimeric forms of AtPRMT10 and PRMT3 show that dimer formation produces coherent motions in key catalytic domains (Figure 2.9 and Supplemental Figure 2.3). PRMT dimers exhibit reduced fluctuations in the N-terminal αY-loop-αZ region, which not only forms direct contacts with the SAM methyl donor, but also forms a portion of substrate binding groove I that is conserved among PRMTs. Furthermore, dimerization results in more correlated motions throughout the SAM binding domain. Previous studies have shown that oligomerization can facilitate protein-ligand interaction by increasing the correlation in the motion of the structural elements involved in ligand binding (*35*). Importantly, our results show that the effects of dimerization on the motion of AtPRMT10 can be generalized into other members of the PRMT family (Supplemental

Figure 2.3). Dimerization appears to facilitate the methyltransferase activity of PRMTs by producing coherent protein motions in the SAM binding region.

Members of the PRMT family have a relatively conserved catalytic core, but exhibit remarkable diversity in the length and sequence of their N-terminal regions. Multiple lines of evidence suggest that the variations in the N-terminus diversify the functions of the PRMT family by modulating the substrate specificities (*17, 36-38*). AtPRMT10 has a 30-residue N-terminal addition, which is one of the shortest among known PRMTs. Secondary structure analysis predicts that the N-terminal addition of AtPRMT10 remains in a disordered state. In support of this prediction, the AtPRMT10 N-terminal addition is prone to proteolysis (data not shown), and is not ordered in our crystal structure. Although PRMT1 also has a short N-terminal region (~30 residues), its length varies more among different PRMT1 isoforms and these variations have been shown to alter the substrate specificity of PRMT1 (*39*).

The results presented here indicate that residues 1-10 can affect the substrate specificity of AtPRMT10 (Figure 2.5a, b). The deletion of the N-terminal addition enhances the activity of AtPRMT10 toward histone H2A, but does not significantly alter AtPRMT10 activity toward histone H4. This variation may result from the difference in the way that H4 and H2A interact with AtPRMT10. Based on the crystal structure of dimeric AtPRMT10, the 30-residue N-terminal addition is likely located at one side of the ring, adjacent to substrate binding grooves III and IV, but distant from substrate binding grooves I and II. Thus, H2A may employ AtPRMT10 substrate binding groove III or IV, while H4 employs substrate binding groove I. The local sequence of the methylation site in H2A (SG***R***$_3$GKGG) is identical to that of H4 (SG***R***$_3$GKGG), indicating that the sequence outside the methylation site is also important for the interaction of PRMT with its substrates. In support of this notion, our results demonstrate that deletion of the sequence C-terminal to residue 20 of H4 dramatically reduced the methylation at arginine 3 by AtPRMT10 (Figure 2.5c).

PRMT10 displayed comparable activities toward histone H4 and N-terminally GST-tagged histone H4 (GST-H4) (Figure 2.5c). In histone H4, the methylation site arginine-3 is located proximal to the N-terminus. Therefore, the standard H4 substrate can bind to the AtPRMT10 binding grooves in a linear fashion, as modeled in the crystal structure of PRMT1-peptide complex (*20*). Our results with the GST-H4 substrate, however, indicate that the target arginine of AtPRMT10 could be located internally within a larger protein, rather than only within an N-terminal tail. Such an observation expands the scope of potential substrates for AtPRMT10 to include proteins with target arginines located on flexible central loops capable of accessing the AtPRMT10 active site. Such substrate proteins may be identified in *Arabidopsis* that impact flowering time in a AtPRMT10-dependent manner. In summary, the data presented here indicate that, while the PRMTs share some key traits (*e.g.,* a functional dimer and coherent SAM-binding domain motion), unique features of specific PRMTs, like the larger central cavity of the AtPRMT10 dimer, may lead to unique methylation patterns and target substrate proteins.

2.4 METHODS

Cloning, expression and purification of AtPRMT10

The expression plasmids encoding wild-type AtPRMT10 (1-383) and its various mutants and related constructs were created using the standard ligation-independent cloning techniques, as described by Stols *et al.* (*40*). All expression plasmids used in this study were sequence verified. AtPRMT10 was overexpressed in *Escherichia coli* BL21-CondonPlus (DE3) RIPL (Stratagene) and purified as described previously with some modifications (*41*). The cells were grown at 37 ^{o}C to an OD_{600} of 0.6 in Luria-Bertani media containing 50 μg/mL chloramphenicol and 50 μg/mL ampicilin. Protein expression was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final

concentration of 0.1 mM and the culture was grown for another 16 h at 18 °C. The harvested cells were resuspended in buffer A (50 mM Na phosphate pH 7.4, 500 mM NaCl and 20 mM imidazole) supplemented with 0.5 mM EDTA, 0.1% Triton X-100, 1 mM phenylmethylsulphonyl fluoride (PMSF), one tablet of a protease inhibitor cocktail (Roche), and 1 mg/mL lysozyme. After 45 min of incubation on ice, the resuspended cells were sonicated on ice for 3 min and the lysate was centrifuged at 50,000 × g for 60 min at 4 °C. The supernatant was passed through a 0.2 µm filter (Millipore) and then loaded onto a 5 mL high performance HisTrap™ column (GE Life Sciences), equilibrated with buffer A. The column was washed with 100 mL buffer A to remove nonspecifically bound proteins; the bound protein was then eluted with buffer B (50 mM Na phosphate pH 7.4, 50 mM NaCl and 250 mM imidazole). The elutent was loaded onto a HiPrep™ 26/10 desalting column (GE Healthcare Life Sciences) equilibrated with buffer C (20 mM Tris-HCl pH 8.0 and 150 mM NaCl), and the protein-containing fractions were collected. To remove the His-MBP tag, TEV protease was added into the pooled protein fractions at a ratio of 1:100 (w/w) TEV to TraI. After 12 hr of incubation at 4 °C, the mixture was reloaded onto the 5 mL high performance HisTrap™ column (GE Life Sciences), equilibrated with buffer A. The flow-through fractions were collected and concentrated in a Centricon YM10 (Amicon) concentrator. Concentrated protein was loaded on a HiLoad™ 16/60 Superdex 200 column (GE, Life Sciences) equilibrated with sizing buffer (20 mM Tris-HCl pH 7.5, 250 mM NaCl, 5% Glycerol). AtPRMT10 containing fractions were concentrated, flash frozen in liquid nitrogen and stored at -80 °C. Purified protein was >95% pure by SDS-PAGE.

Crystallization, data collection, structure determination, and refinement

Diffraction-quality crystals of AtPRMT10 (residues 11-383)-SAH complex were obtained by the hanging-drop vapour-diffusion method at 22 °C, with the mother liquid solution containing 0.1 M Tris-HCl pH 7.6, 2.3 M Na_2HPO_4 and 0.1 M arginine. Crystals grew

to the size of 250 × 200 × 50 μm in approximately 10 days. Since flash-frozen crystals diffracted poorly and could not be used for structural determination, diffraction data were collected from warm-mounted crystals to 2.6 Å resolution using a Rigaku X-ray generator MicroMax-007HF. Data from four different crystals were reduced and merged using the program HKL2000 (*42*) (Table 2.1). Data quality was examined using the program PHENIX (*30*). The structure was determined in space group $P2_1$ by molecular replacement using the program PHENIX (*30*). The crystal structure of rat PRMT3 (PDB entry, 1F3L), processed using the program chainsaw of the CCP4 package (*43*), was used as the template for molecular replacement. Due to the salient difference between AtPRMT10 and PRMT3 in the sequence of the dimerization arm, the dimerization arm of PRMT3 (residues 370-399) was not incorporated into the template. Since psudomerohedral twinning (approximately 50%) was detected with the crystals used for the structure determination, least square twin refinement was performed using the program PHENIX. The structural model was further built manually using the program Coot (*44*), and refined using the program PHENIX.

Methyltransferase activity assay

In vitro methyltransferase assays were performed as described previously (*29*). In brief, the reaction mixture contained 20 mM Tris-HCl pH 8.0, 4 mM EDTA, 1 mM DTT, 0.5 mM PMSF, 4 μM *S*-[Methyl-^{3}H]-Adenosyl-*L*-methionine (Perkin Elmer [NET155]) and indicated concentrations of AtPRMT10 and protein substrates. The reaction mixtures were separated on a 15% SDS-PAGE and stained with commassie blue. The gel was then treated with Amplifier (GE Healthcare Life Sciences), dried and exposed to Kodak Biomax MS film at -80 ^{o}C.

Dynamic light scattering (DLS)

The hydrodynamic radii of various AtPRMT10 constructs were measured by a DynoPro DLS system (Wyatt Technology Corporation). All samples and buffers (20 mM Tris-

HCl pH 7.5, 100 mM NaCl, 0.5 mM EDTA) were filtered through 0.2 μM filters (Millipore) or centrifuged at 17,000 × g at 4 ^{o}C for 30 min before measurement. Three replicates were performed for each sample. The hydrodynamic radii and molecular weights of AtPRMT10 samples were estimated using the assumption of globular protein shape.

Molecular dynamics (MD) simulations

MD simulations of AtPRMT10 were performed using the AMBER 2003 force field (*45*) as described previously (*35*). All production runs were generated using the PMEMD module of Amber 9.0 (*46*) with a 2 fs time step. The topology and parameter files were created using the LEaP program within AMBER (*46*). To maintain charge neutrality, the protein molecule was surrounded by a truncated octahedron of water and sodium ions in the simulation system. Electrostatic interactions were calculated using the particle-mesh Ewald algorithm (*47*) with a cutoff of 10 Å applied to Lennard-Jones interactions. Energy minimization was conducted using the SANDER package within AMBER (*46*). Equilibration consisted of 20 ps of constant volume conditions with heating from 100 to 300 K and subsequent 100 ps of constant temperature conditions.

Simulation results were analyzed by the PTRAJ package in AMBER (*46*). The pair-wise correlation coefficient, C_{ij}, was calculated between the α-carbons of two residues as described by Sharma *et al*. (*48*). When the two residues i and j move in a correlated fashion (the angle between the motion of i and j is less than 90 º), $0<C_{ij}\leq1$; when they move in an anti-correlated way (the angle between the motion of i and j is more than 90 º but less than 180 º), $-1\leq C_{ij}<0$; finally, when they move in a non-correlated manner (randomly), $C_{ij}=0$. The more positive the value of C_{ij} is, the smaller the angle between the motion of the two residues is. Single-linkage clustering analysis was performed to identify groups of residues that move in a correlated or anti-correlated fashion, as described by Leese *et al*. (*49*).

2.5 ACKNOWLEDGEMENTS

We thank Dr. Michael Miley at the UNC Biomolecular X-ray Crystallography Facility for assistance in sample preparation and data collection. Also, we thank Dr. Feng Ding for assistance in simulation analysis. This work was supported by NIH grant AI78924, the National Basic Research Program of China (grant no.2009CB941500), and the National Natural Science Foundation of China (grant no. 30921061).

2.6 FIGURE LEGENDS

Figure 2.1 Crystal structure of AtPRMT10. (a) Domain architecture of AtPRMT10 from *Arabidopsis thaliana*. The SAM binding domain (residues 31-174) is shown in red, the β-barrel domain (residues 175-186, 237-383) in blue, and the arm domain (or dimerization domain, residues 187-236) in yellow. **(b)** The crystal structure of the AtPRMT10-SAH complex with key secondary structure elements labeled (helices are labeled X through I and β-strands are numbered 1 through 15). The bound SAH is shown as sticks and spheres. The structure is colored as in Figures 2.1A-B. The first and last residues of AtPRMT10 are indicated. **(c)** Two views of the superimposition of AtPRMT10 (residues 31-383, magenta) with rat PRMT1 (residues 41-353, cyan, PDB entry 1ORI). Key structural differences between AtPRMT10 and PRMT1 (located in the dimerization domain and two loops in the β-barrel domain) are indicated by arrows. **(d)** A stereo-view representation of SAH binding. A 2.6 Å resolution simulated annealing omit map of SAH contoured at 2.5σ (blue mesh) is shown. Hydrogen bonds are indicated by red dashed lines.

Figure 2.2 Structure-based sequence alignment of AtPRMT10, rat PRMT1 (PDB: 1ORI), rat PRMT3 (PDB: 1F3L), yeast RMT1 (1G6Q) and mouse CARM1 (PDB: 3B3F). Secondary-structure elements are shown across the top of the aligned sequences. Residue

numbers are provided on the right. Invariant and similar residues are highlighted in black and gray, respectively. Protein domains are colored as in Figures 2.1A-B. The four PRMT signature motifs are labeled. Residues involved in SAM binding and dimerization are highlighted by red and black stars, respectively.

Figure 2.3 Dimer formation of AtPRMT10. (a) The crystal structure of an AtPRMT10 dimer. The structure is colored and labeled as in Figures 2.1A-B. The location of each dimer interface is highlighted by an arrow. The dimer is formed by the interaction between the dimerization arm of one monomer and the outer surface of the SAM binding domain of the other monomer. **(b)** A surface representation of two views of an AtPRMT10 dimer, with two monomers are colored in gray and pink, respectively. **(c)** An expanded stereo-view of the dimer interface. Two monomers are colored as in Figure 2.2B. Residues involved in dimer formation are shown as sticks and labeled. Hydrogen bonds are indicated by red dashed lines.

Figure 2.4 Surface representation of various PRMT paralogs, including rat PRMT1 (PDB: 1ORI) (a), rat PRMT3 (PDB: 1F3L) **(b)**, mouse CARM1 (PDB: 3B3F) **(c)** and AtPRMT10 **(d)**. For consistency, the sequences N-terminal to helix X (including helix X) are deleted from the structure. The dimensions of the central cavities are indicated.

Figure 2.5 Methyltransferase activities of different AtPRMT10 constructs *in vitro*. (a) Indicated AtPRMT10 proteins (5 µg) were used for *in vitro* methylation activity assay. The reaction mixture was separated on a 15% SDS-PAGE and the autoradiograph of the gel is recorded. The experiments were performed in triplicate and a typical result is shown here. **(b)** Quantification of the results from Figure 2.5A. The relative activities presented here were calculated by considering the activity of wild-type AtPRMT10 over H2A as one. The activities

of AtPRMT10 mutants over H2A are significantly higher than that of wild-type enzyme (n=3; error bars represent SEM; Student t test *p < 0.03, **p < 0.002). **(c)** Methyltransferase activities of different AtPRMT10 constructs on H4, GST-H4 and H4N1-20.

Figure 2.6 Surface electrostatic potential of AtPRMT10. Acidic surfaces are represented in red; basic surfaces in blue and neutral surfaces in gray. Acidic surface residues are labeled and colored based on their conservation among five PRMT paralogs as indicted by the sequence alignment in Figure 2.2 (black, conservation≥80%; green, 40%<conservation<80%; blue, 20%<conservation≤40%; unique for AtPRMT10, red). The location of putative substrate binding grooves, the active site, the dimer interface and helix αX are highlighted and labeled. The top left view has the same orientation as in Figure 2.1B.

Figure 2.7 AtPRMT10 exhibits a uniquely accessible active site. (a) Superimposition of AtPRMT10 (magenta) with rat PRMT1 (cyan, PDB: 1ORI), rat PRMT3 (yellow, PDB: 1F3L) and mouse CARM1 (blue, PDB: 3B3F). The bottom monomers are used for the alignment. The left view rotates about 90^{o} along the vertical axis with respect to the view shown in Figure 2.3A. The angles of the rotation of AtPRMT10 and CARM1 relative to PRMT3 are shown. For clarity, the bottom monomers are not shown in the middle panel. The left edges of four monomers are indicated by vertical dashed lines with corresponding colors. The distance of the leftward translation of AtPRMT10 and CARM1 relative to PRMT3 are shown. The right panel is a schematic representation of the middle panel, with PRMT monomers shown as rectangles. The superimposed bottom monomers are shown and colored in white. **(b)-(e)** Views of the dimeric AtPRMT10, rat PRMT1, rat PRMT3 and mouse CARM1. The top and bottom monomers are shown in surface and ribbon representation, respectively. The cofactor analog SAH is shown in surface representation. The accessibility angle for the active site of each PRMT is labeled.

Figure 2.8 Conservation of total energy during AtPRMT10 simulations. Total energy, an indicator of the overall simulation stability, remains relatively constants during the course of MD simulations of monomeric PRMT10 **(a)** and dimeric PRMT10 **(b)**. The average of the total energy is shown in a blue lines **(a)** and a green line **(b)**, respectively. The final 10 ns of each simulation, highlighted by red square, was used for analysis.

Figure 2.9 Effects of dimerization on the motion of AtPRMT10. (a) Local fluctuation of residues in dimeric and monomeric AtPRMT10. Major differences between monomeric and dimeric AtPRMT10, including the two N-terminal helices αY-αZ and the dimerization arm (αE-αG), are highlighted by arrows and labeled. **(b)** The structure of a dimeric AtPRMT10. The two monomers are colored in blue and green respectively. The regions that displayed dramatically reduced local fluctuations in dimeric PMT10 are colored in light green in one monomer and in light blue in the other monomer. The putative substrate sites are represented in orange balls. Covariance analysis of dimeric AtPRMT10 **(c)** and monomeric AtPRMT10 **(d).** The values of residue-residue correlation coefficients range from blue (anticorrelated, -0.62) to red (correlated, +1.0), with non-correlated residue pairs colored in yellow. Schematic representations of the secondary structure corresponding to the residues on x-axis and y-axis are presented from left to right and bottom to top. **(e)**, **(f)** Clustering of correlated residues in dimeric AtPRMT10 **(e)** and monomeric AtPRMT10 **(f)**. Clusters with correlation coefficients higher than 0.7 are shown in different colors (other than gray) and all other regions are colored in gray. SAH is shown as sticks and spheres to highlight the location of the SAM binding pocket, although it is removed from the crystal structures before MD simulations. **(g)** An expanded stereo-view of the SAM binding domain of Figure 2.9E, with the secondary structures labeled.

2.7 SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 2.1 Sequence alignment of AtPRMT10 orthologs in various plants. The aligned sequences include AtPRMT10 from *A. thaliana* (GeneBank accession number NP562720), *G. max* (ACU22946), *V. vinifera* (XP002285026), *P. trichocarpa* (XP002332378), *Z. mays* (NP001170456), *S. bicolor* (XP002436453) and *O. sativa* (EAY99613). The alignment was conducted using Clustal W, and then manually adjusted. The secondary-structure elements and residue numbering of AtPRMT10 are shown across the top of the sequences, colored and labeled as shown in Figure 2.1b. Residues involved in SAH binding and dimer formation are highlighted by red and black stars, respectively. Four PRMT signature motifs are labeled.

Supplemental Figure 2.2 Conserved residues between AtPRMT10 and its paralogs, including rat PRMT1, rat PRMT3, yeast RMT1 and mouse CARM1, were mapped onto the structure of AtPRMT10. All three views are shown in surface representation where magenta represents residues with 100% conservation and orange denotes residues with high similarity. The top left view has the same orientation as shown in Figure 2.1b. The location of dimer interface and active sites are labeled.

Supplemental Figure 2.3 Effects of dimerization on the motion of rat PRMT3. (a) Local fluctuation (as determined by B factors) of residues in dimeric and monomeric PRMT3 (PDB: 1F3L). Major differences between monomeric and dimeric PRMT3, including the two N-terminal helices αX-αY-αZ and the arm (αE-αF-αG) are highlighted by arrows and labeled. Schematic representations of the secondary structure corresponding to the residues on x-axis are presented from left to right and bottom to top. Covariance analysis of dimeric PRMT3 (b) and monomeric PRMT3 (c). The values of residue-residue correlation

coefficients range from blue (anticorrelated, -0.62) to red (correlated, +1.00), with non-correlated residue pairs colored in yellow. Schematic representations of the secondary structure corresponding to the residues on x-axis and y-axis are presented from left to right and bottom to top.

Table 2.1 Crystallographic data and refinement statistics

Resolution (Å) (highest shell)	33.0–2.61 (2.67–2.61)
Space group	$P2_1$
Unit cell parameters	
a, b, c (Å)	80.55, 86.69, 114.74
β (°)	89.97
Twinning fraction	0.5
Number of total reflections	48,404
Number of unique reflections	46,744
R_{sym}[a] (%) (highest shell)	12.0 (58.0)
Completeness (%) (highest shell)	96.6 (91.0)
Mean I/σ (highest shell)	13.0 (2.30)
Average redundancy	2.90 (2.60)
R_{cryst}[b] (%) (highest shell)	18.2 (35.0)
R_{free}[c] (%) (highest shell)	22.4 (41.2)
RMSD	
Bond lengths (Å)	0.007
Bond angles (°)	1.17
Dihedral angles (°)	16.2
Number of atoms per asymmetric unit	
Protein	10,751
Ligand	104
Solvent	174

a $Rsym=\sum|I-\langle I\rangle|/\sum I$, where I is the observed intensity and $\langle I\rangle$ is the average intensity of multiple symmetry-related observations of the reflection.

b $Rcryst=\sum||Fobs|-|Fcalc||/\sum|Fobs|$, where F_{obs} and F_{calc} are the observed and calculated structure factors, respectively.

c $Rfree=\sum||Fobs|-|Fcalc||/\sum|Fobs|$ for 5% of the data not used at any stage of refinement.

Table 2.2 Oligomeric states of AtPRMT10 mutants

	Monomer size (kDa)	Gel-filtration size (– SAH) (kDa)	– SAH			+ SAH (1 mM)	
			Gel-filtration size/ monomer size	DLS Size (kDa)	DLS size/ monomer size	DLS size (kDa)	DLS size/ monomer size
Wild type	43.1	80	1.8	93	2.2	92	2.1
ΔN10	42.1	ND	ND	89	2.1	87	2.1
ΔN20	41.4	ND	ND	83	2.0	83	2.0
ΔN30	40.4	68	1.7	74	1.8	80	2.0
Δ203 –225	40.8	34	0.9	66	1.6	61	1.5

ND, not determined.

Figure 2.1

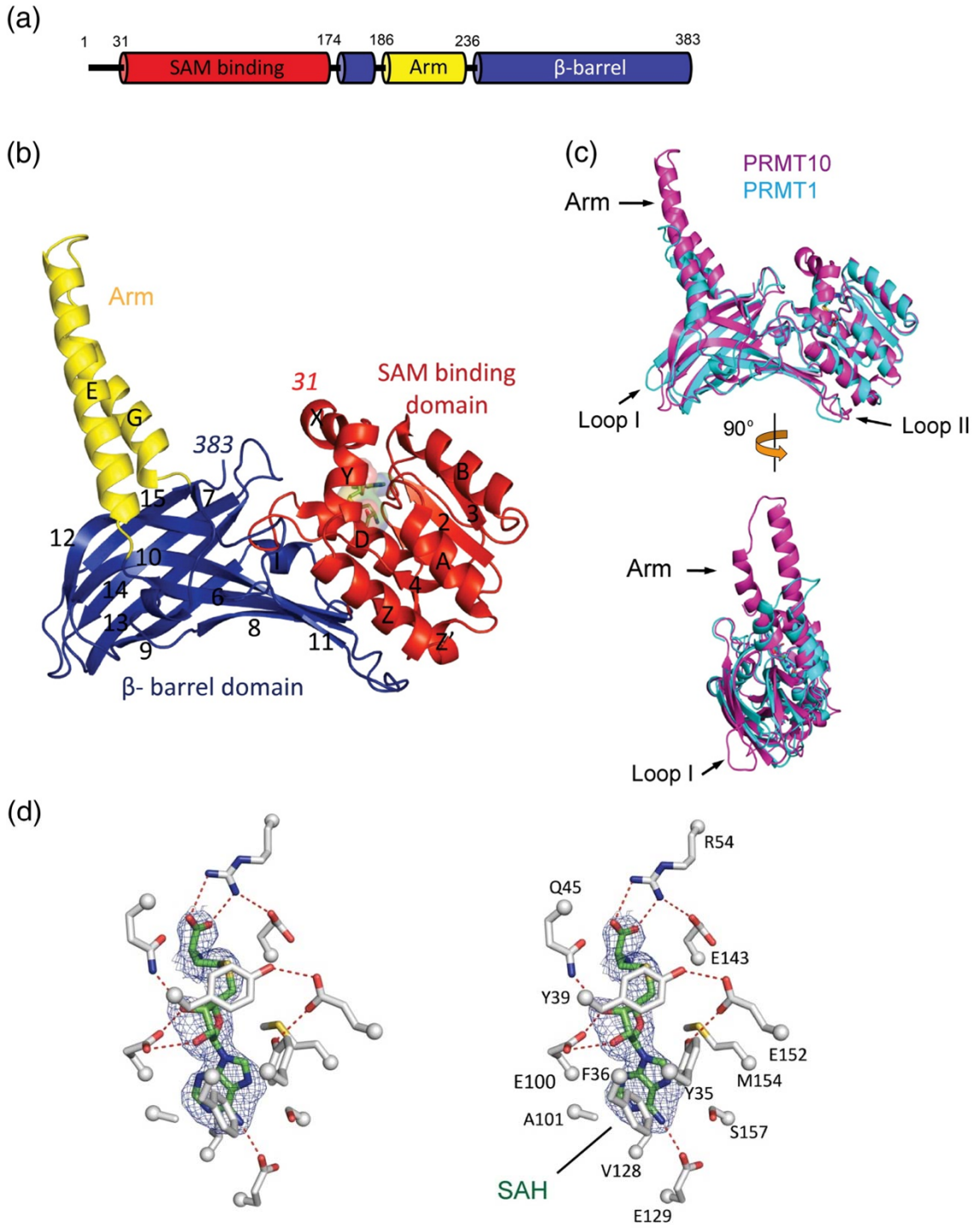

(a)
1
31
174
186
236
383
SAM binding
Arm
β-barrel
(b)
Arm
E
G
31
383
SAM binding domain
X
Y
B
2
3
D
A
4
Z
Z'
1
15
7
12
10
14
13
6
8
11
9
β- barrel domain
(c)
PRMT10
PRMT1
Arm
Loop I
90°
Loop II
Arm
Loop I
(d)
R54
Q45
E143
Y39
E152
E100
F36
M154
Y35
A101
S157
V128
SAH
E129

Figure 2.2

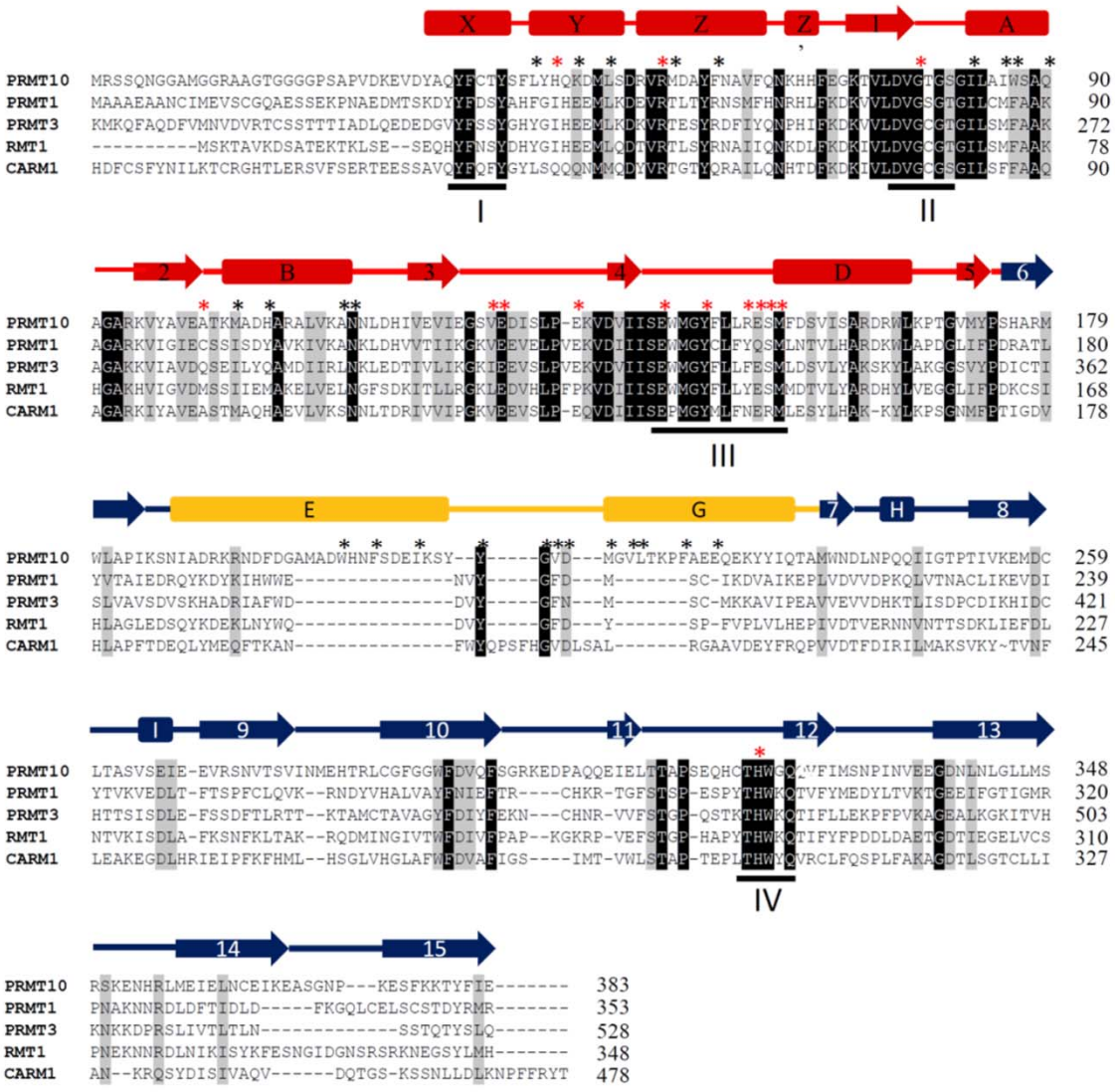

X
Y
Z
Z'
1
A
PRMT10 MRSSQNGGAMGGRAAGTGGGGPSAPVDKEVDYAQYFCTYSFLYHQKDMLSDRVRMDAYFNAVFQNKHHFEGKTVLDVGTGSGILAIWSAQ 90
PRMT1 MAAAEAANCIMEVSCGQAESSEKPNAEDMTSKDYYFDSYAHFGIHEEMLKDEVRTLTYRNSMFHNRHLFKDKVVLDVGSGTGILCMFAAK 90
PRMT3 KMKQFAQDFVMNVDVRTCSSTTTIADLQEDEDGVYFSSYGHYGIHEEMLKDKVRTESYRDFIYQNPHIFKDKVVLDVGCGTGILSMFAAK 272
RMT1 ----------MSKTAVKDSATEKTKLSE--SEQHYFNSYDHYGIHEEMLQDIVRTLSYRNAIIQNKDLFKDKIVLDVGCGTGILSMFAAK 78
CARM1 HDFCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMQDVVRTGTYQRAILQNHTDFKDKIVLDVGCGSGILSFFAAQ 90
I
II
2
B
3
4
D
5
6
PRMT10 AGARKVYAVEATKMADHARALVKANNLDHIVEVIEGSVEDISLP-EKVDVIISEWMGYFLLRESMFDSVISARDRWLKPTGVMYPSHARM 179
PRMT1 AGARKVIGIECSSISDYAVKIVKANKLDHVVTIIKGKVEEVELPVEKVDIIISEWMGYCLFYQSMLNTVLHARDKWLAPDGLIFPDRATL 180
PRMT3 AGAKKVIAVDQSEILYQAMDIIRLNKLEDTIVLIKGKIEEVSLPVEKVDVIISEWMGYFLLFESMLDSVLYAKSKYLAKGGSVYPDICTI 362
RMT1 HGAKHVIGVDMSSIIEMAKELVELNGFSDKITLLRGKLEDVHLPFPKVDIIISEPMGYFLLYESMMDTVLYARDHYLVEGGLIFPDKCSI 168
CARM1 AGARKIYAVEASTMAQHAEVLVKSNNLTDRIVVIPGKVEEVSLP-EQVDIIISEPMGYMLFNERMLESYLHAK-KYLKPSGNMFPTIGDV 178
III
E
G
7
H
8
PRMT10 WLAPIKSNIADRKRNDFDGAMADWHNFSDEIKSY--Y-----GVD---MGVLTKPFAEEQEKYYIQTAMWNDLNPQQIIGTPTIVKEMDC 259
PRMT1 YVTAIEDRQYKDYKIHWWE---------------NVY-----GFD---M-------SC-IKDVAIKEPLVDVVDPKQLVTNACLIKEVDI 239
PRMT3 SLVAVSDVSKHADRIAFWD---------------DVY-----GFN---M-------SC-MKKAVIPEAVVEVVDHKTLISDPCDIKHIDC 421
RMT1 HLAGLEDSQYKDEKLNYWQ---------------DVY-----GFD---Y-------SP-FVPLVLHEPIVDTVERNNVNTTSDKLIEFDL 227
CARM1 HLAPFTDEQLYMEQFTKAN---------------FWYQPSFHGVDLSAL-------RGAAVDEYFRQPVVDTFDIRILMAKSVKY~TVNF 245
I
9
10
11
12
13
PRMT10 LTASVSEIE-EVRSNVTSVINMEHTRLCGFGGWFDVQFSGRKEDPAQQEIELTTAFSEQHCTHWSQLVFIMSNPINVEEGDNLNLGLLMS 348
PRMT1 YTVKVEDLT-FTSPFCLQVK--RNDYVHALVAYFNIEFTR----CHKR-TGFSTSP-ESPYTHWKQTVFYMEDYLTVKTGEEIFGTIGMR 320
PRMT3 HTTSISDLE-FSSDFTLRTT--KTAMCTAVAGYFDIYFEKN---CHNR-VVFSTGP-QSTKTHWKQTIFLLEKPFPVKAGEALKGKITVH 503
RMT1 NTVKISDLA-FKSNFKLTAK--RQDMINGIVTWFDIVFPAP--KGKRP-VEFSTGP-HAPYTHWKQTIFYFPDDLDAETGDTIEGELVCS 310
CARM1 LEAKEGDLHRIEIPFKFHML--HSGLVHGLAFWFDVAFIGS----IMT-VWLSTAP-TEPLTHWYQVRCLFQSPLFAKAGDTLSGTCLLI 327
IV
14
15
PRMT10 RSKENHRLMEIELNCEIKEASGNP---KESFKKTYFIE------- 383
PRMT1 PNAKNNRDLDFTIDLD-----FKGQLCELSCSTDYRMR------- 353
PRMT3 KNKKDPRSLIVTLTLN-------------SSTQTYSLQ------- 528
RMT1 PNEKNNRDLNIKISYKFESNGIDGNSRSRKNEGSYLMH------- 348
CARM1 AN--KRQSYDISIVAQV------DQTGS-KSSNLLDLKNPFFRYT 478

Figure 2.3

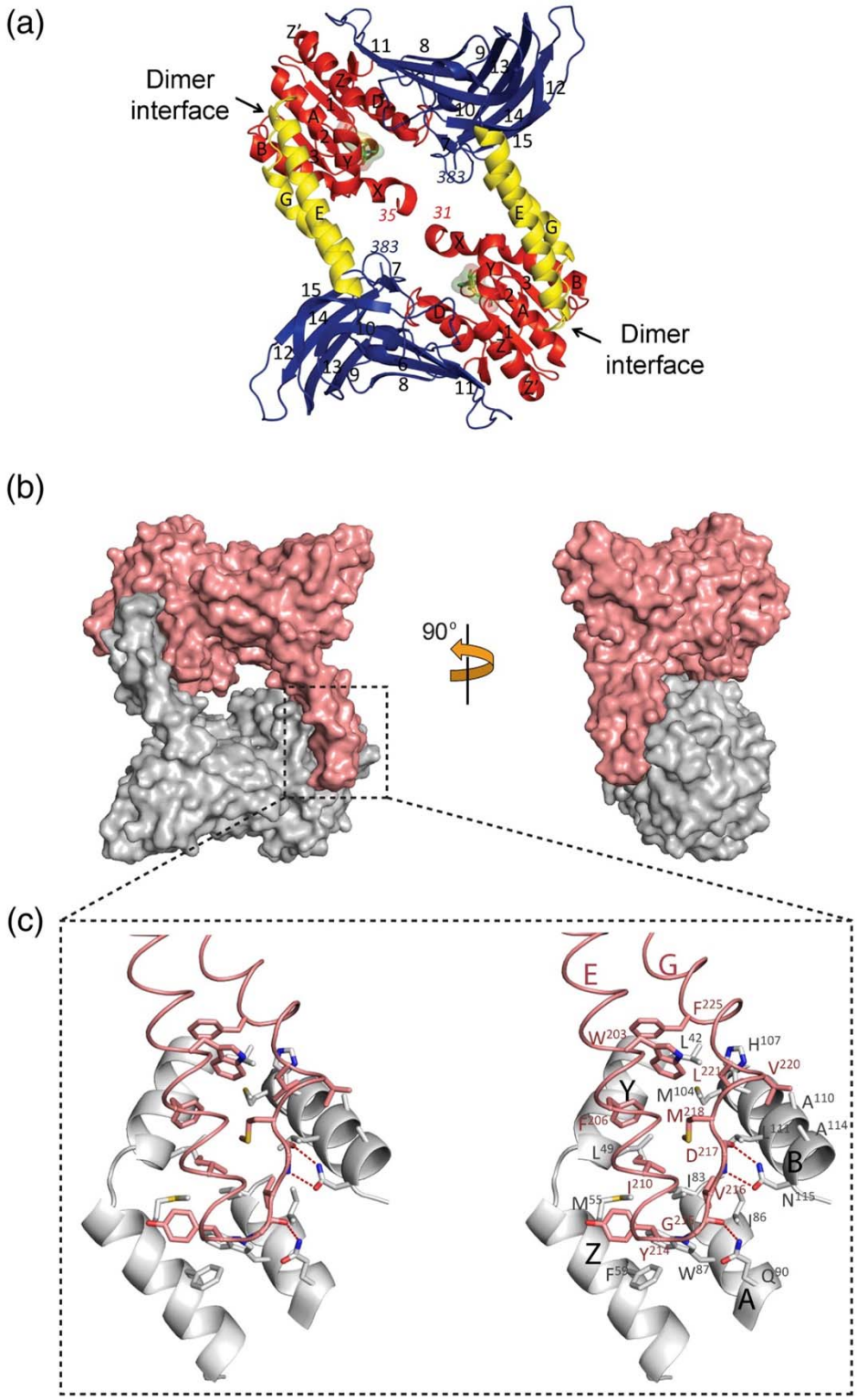


Figure 2.4

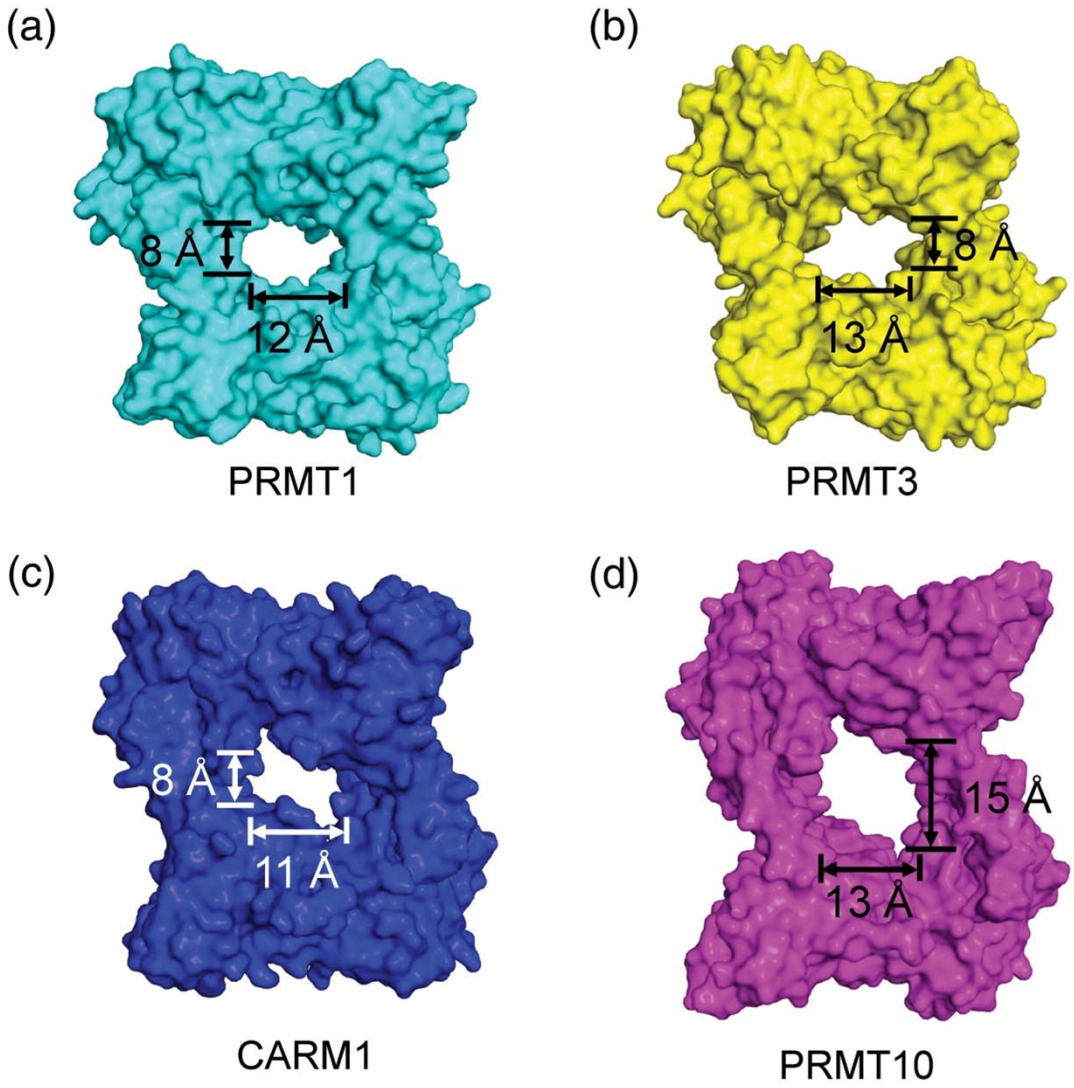


Figure 2.5

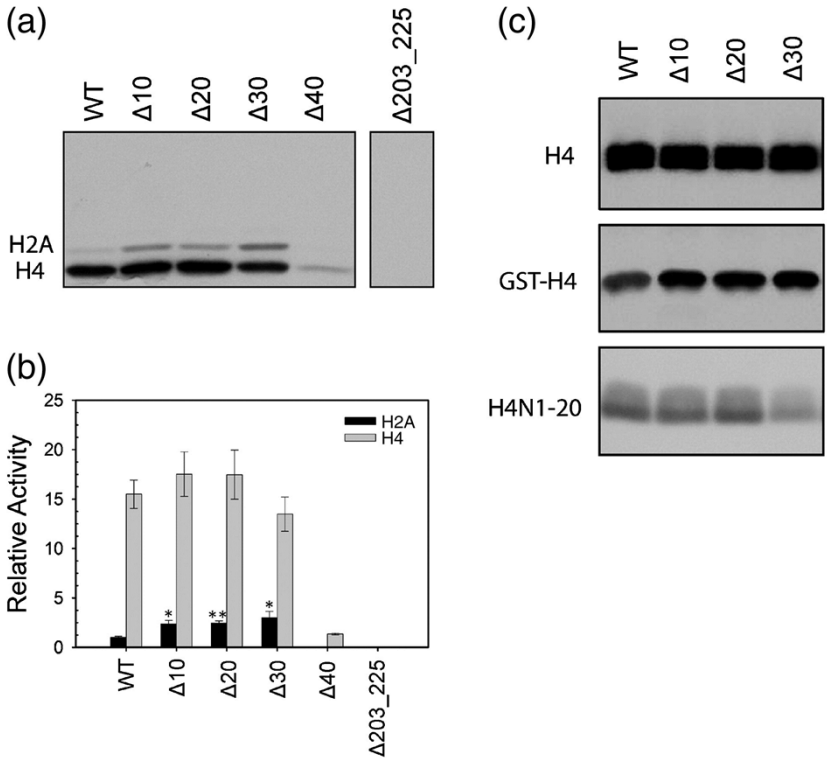

(a)
WT
Δ10
Δ20
Δ30
Δ40
Δ203_225
H2A
H4
(b)
Relative Activity
H2A
H4
25
20
15
10
5
0
*
**
*
WT
Δ10
Δ20
Δ30
Δ40
Δ203_225
(c)
WT
Δ10
Δ20
Δ30
H4
GST-H4
H4N1-20

Figure 2.6

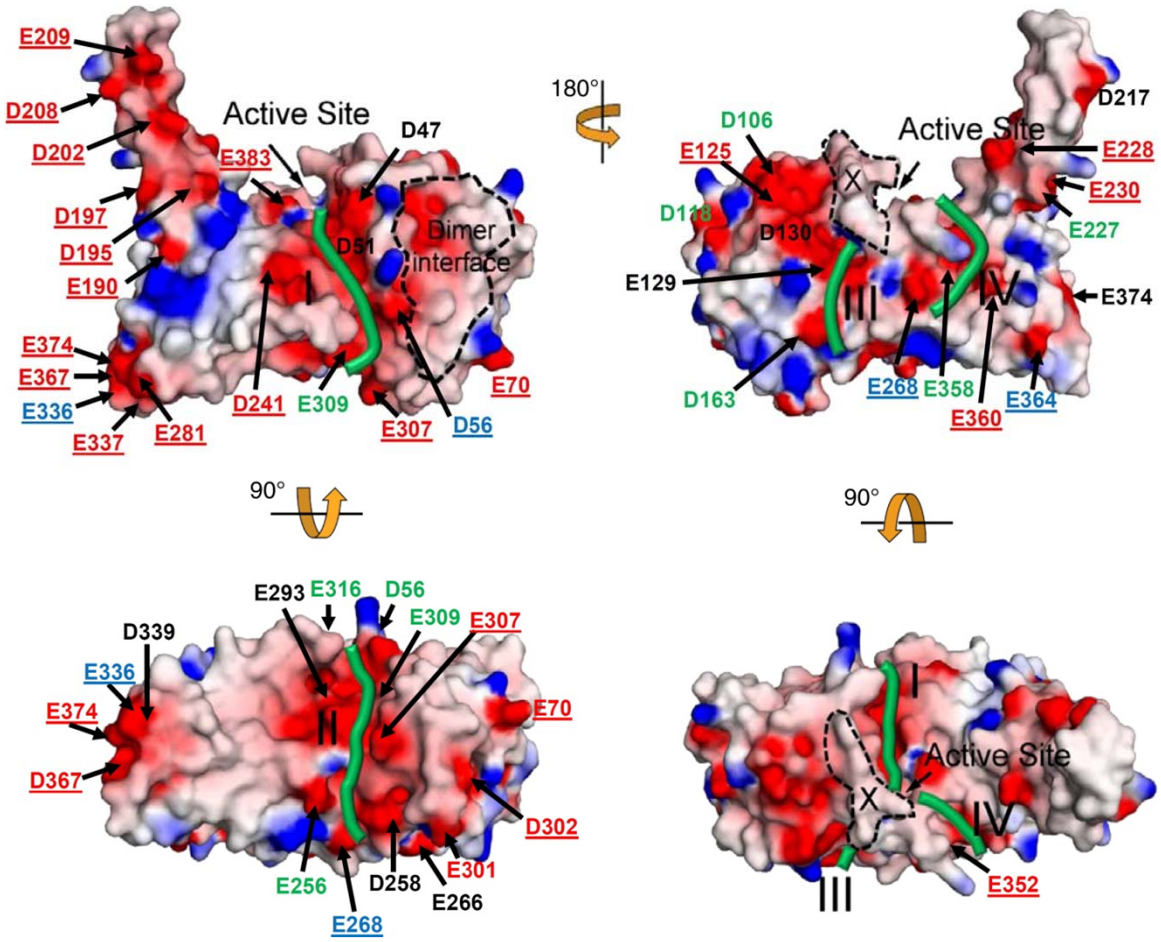

E209
D208
D202
D197
D195
E190
E374
E367
E336
E337
E281
Active Site
D47
E383
D51
Dimer interface
D241
E309
E307
D56
E70
180°
D106
E125
D118
D130
E129
D163
Active Site
D217
E228
E230
E227
E374
E268
E358
E360
E364
90°
90°
E293
E316
D56
E309
E307
D339
E336
E374
D367
E70
D302
E256
D258
E301
E266
E268
Active Site
E352

Figure 2.7

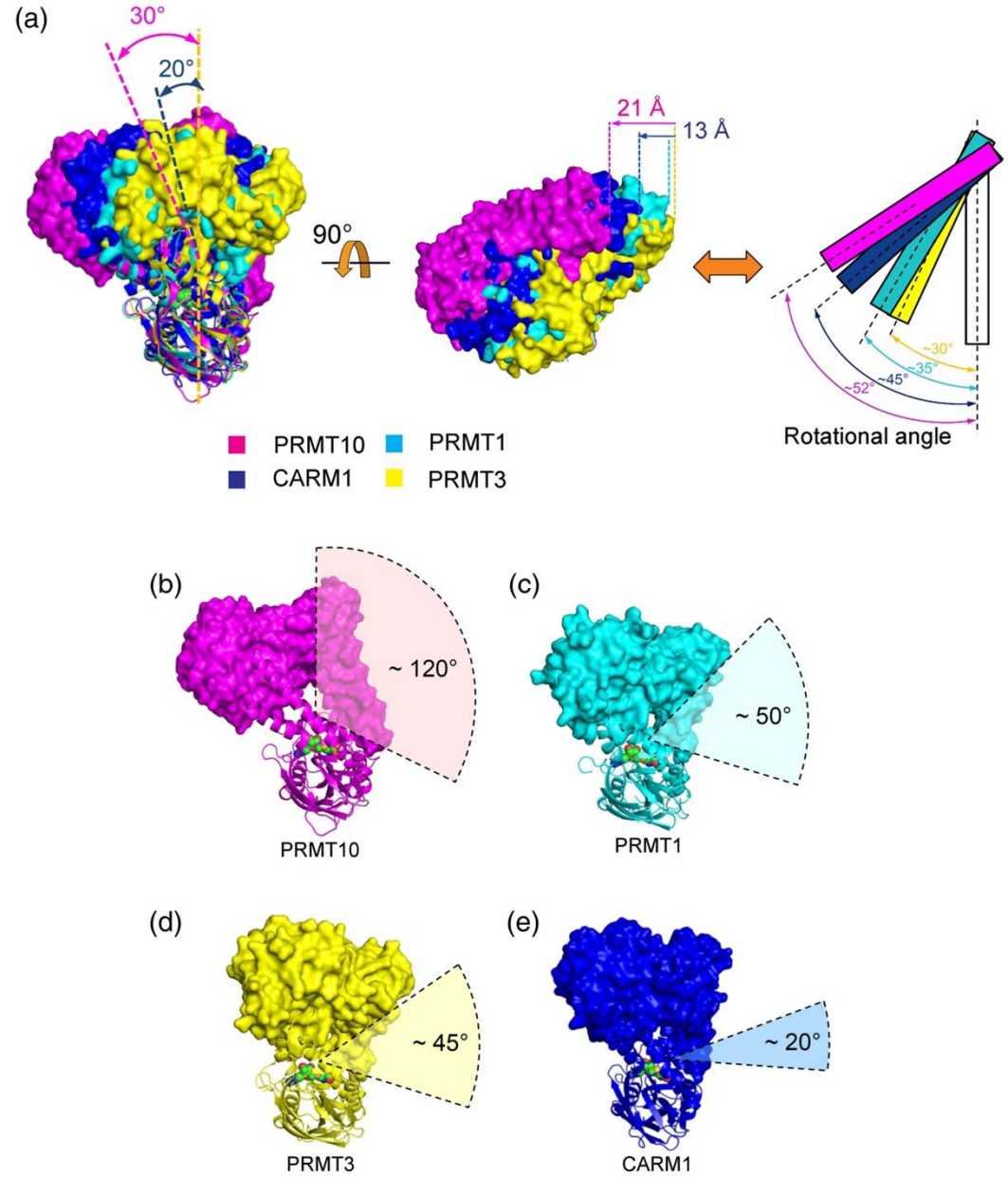

(a)
30°
20°
21 Å
13 Å
90°
~30°
~35°
~45°
~52°
Rotational angle
PRMT10
PRMT1
CARM1
PRMT3
(b)
~ 120°
PRMT10
(c)
~ 50°
PRMT1
(d)
~ 45°
PRMT3
(e)
~ 20°
CARM1

Figure 2.8

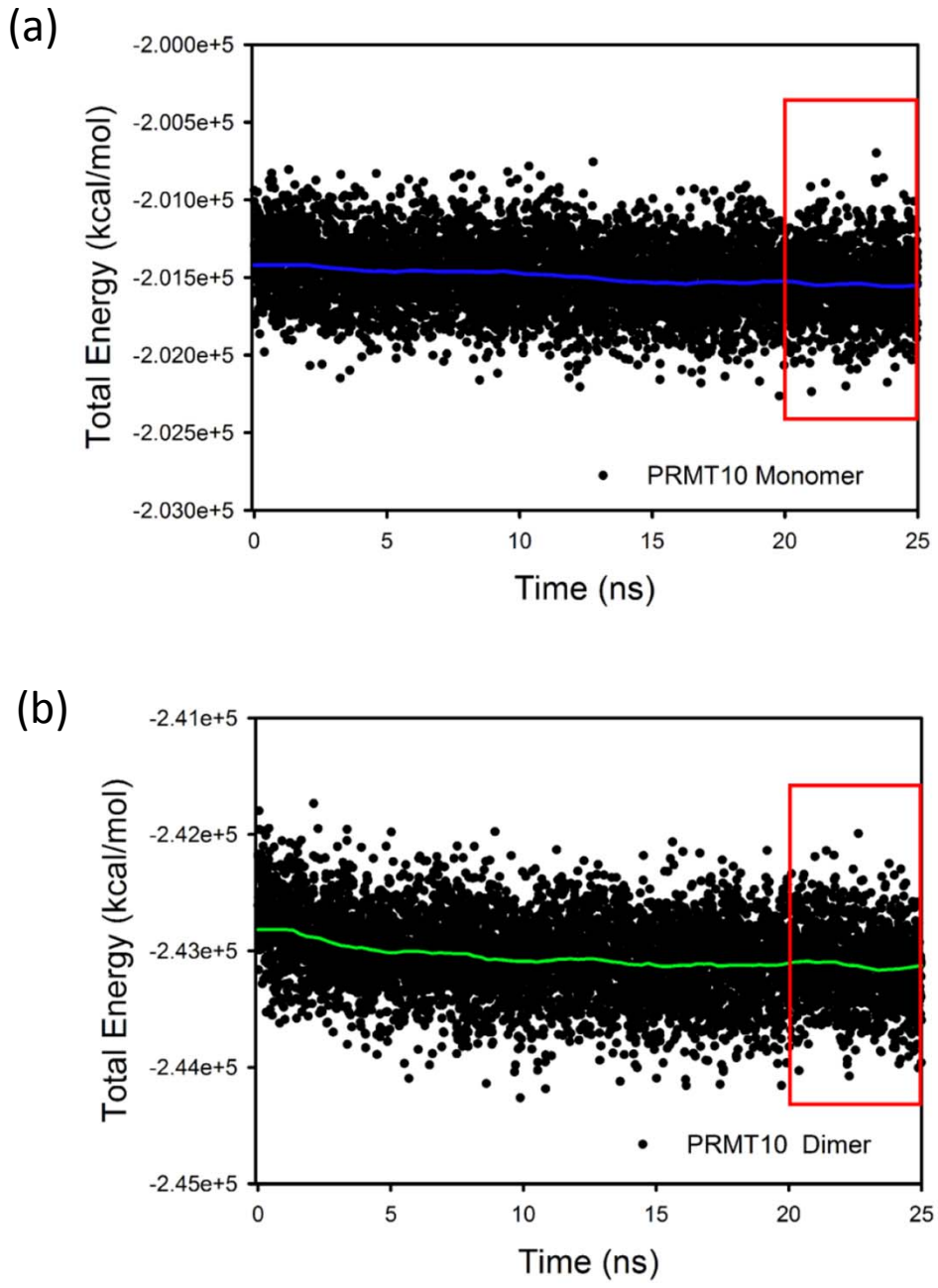
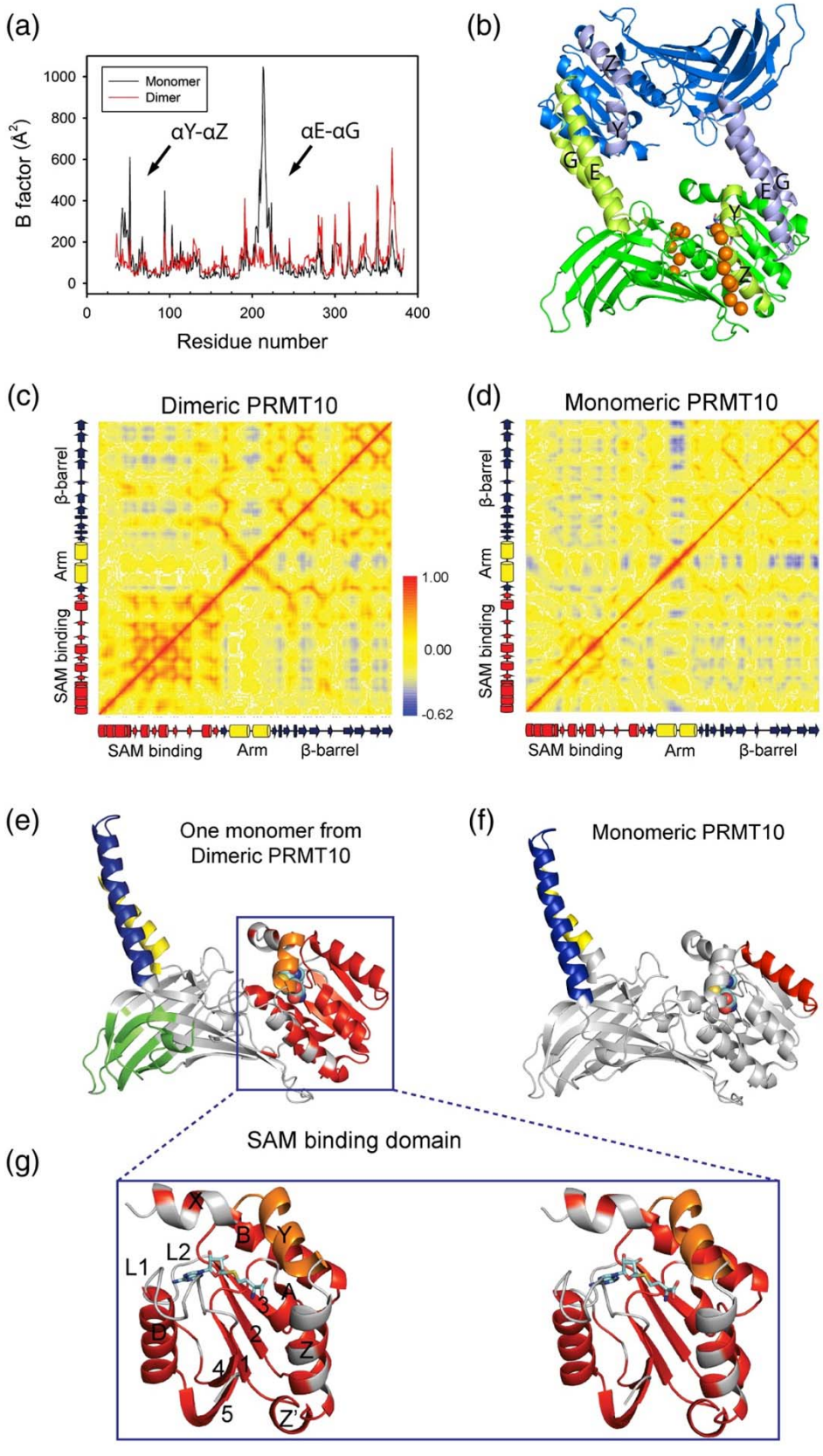

(a)
-2.000e+5
-2.005e+5
-2.010e+5
-2.015e+5
-2.020e+5
-2.025e+5
-2.030e+5
Total Energy (kcal/mol)
PRMT10 Monomer
0
5
10
15
20
25
Time (ns)
(b)
-2.41e+5
-2.42e+5
-2.43e+5
-2.44e+5
-2.45e+5
Total Energy (kcal/mol)
PRMT10 Dimer
0
5
10
15
20
25
Time (ns)

Figure 2.9

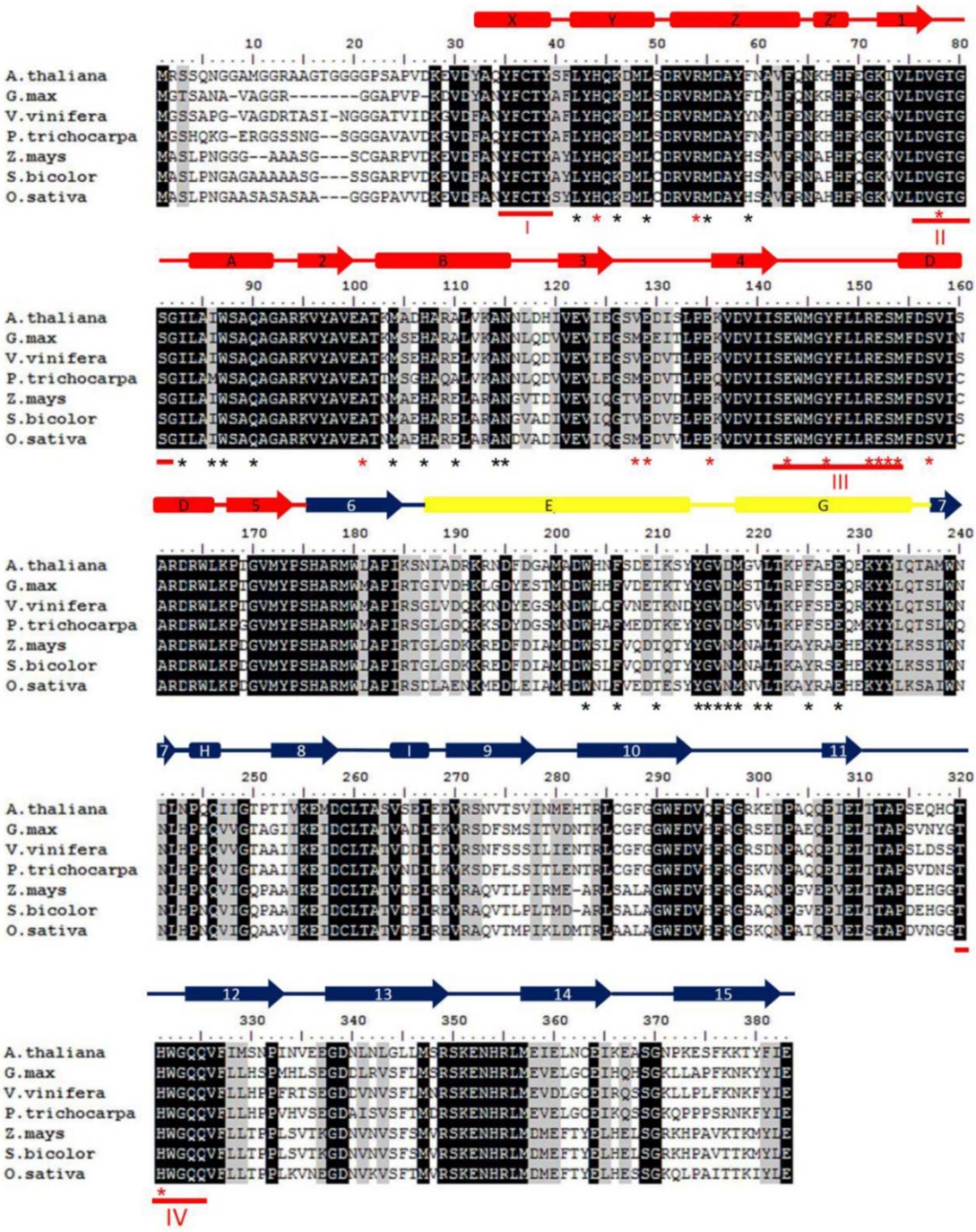

(a)
Monomer
Dimer
αY-αZ
αE-αG
B factor (Å²)
Residue number
(b)
(c)
Dimeric PRMT10
β-barrel
Arm
SAM binding
1.00
0.00
-0.62
(d)
Monomeric PRMT10
(e)
One monomer from
Dimeric PRMT10
(f)
Monomeric PRMT10
(g)
SAM binding domain
L1
L2

Supplemental Table 2.1

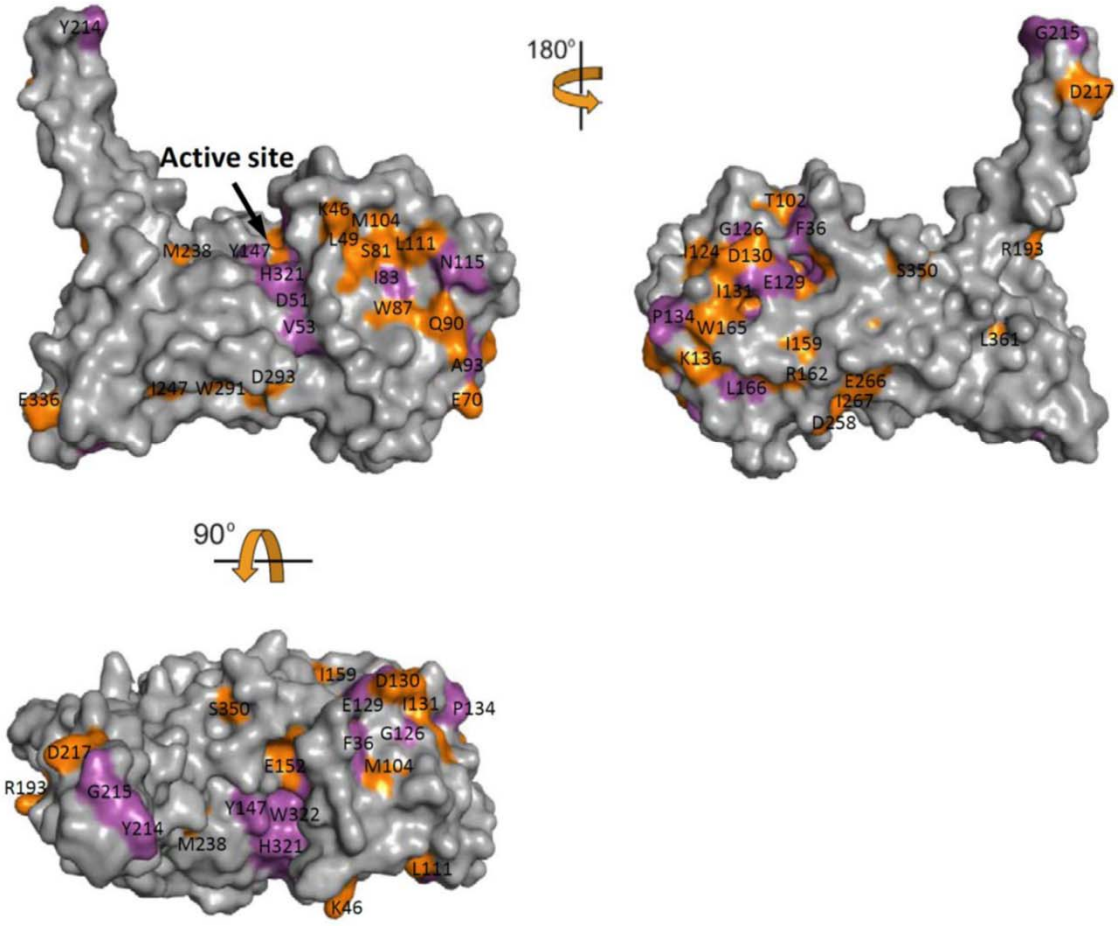
	Twin fraction
H test	0.449
Britton analysis	0.432
Maximum likelihood analysis	0.435

Results from the xtriage analysis in Phenix on the diffraction data used in this study. Three related twin factions were determined.

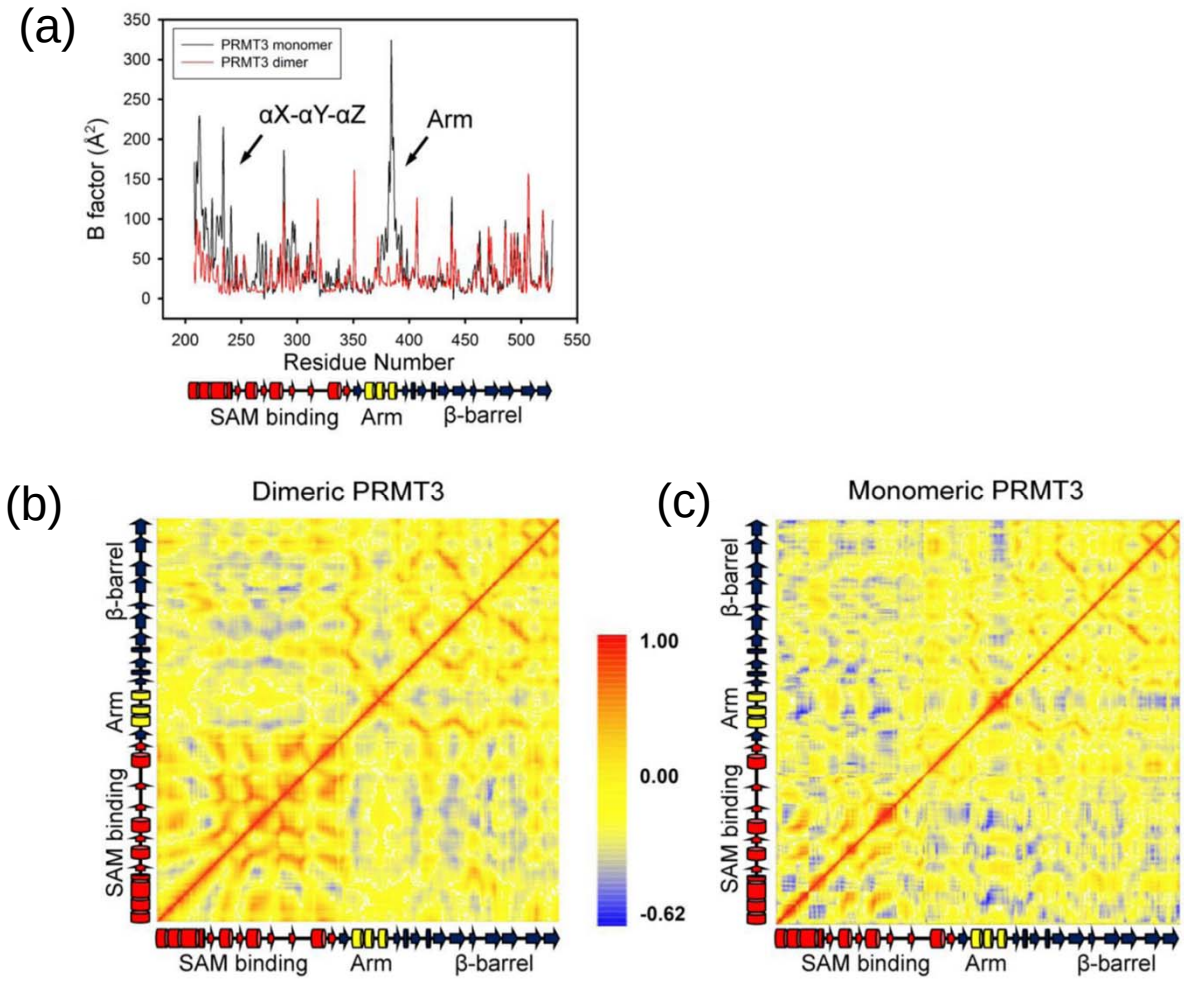
Supplemental Figure 2.1



Supplemental Figure 2.2



Supplemental Figure 2.3



2.9 REFERENCES

1. Bedford, M. T., and Clarke, S. G. (2009) Protein arginine methylation in mammals: who, what, and why, *Mol. Cell. 33*, 1-13.

2. Sanchez, S. E., Petrillo, E., Beckwith, E. J., Zhang, X., Rugnone, M. L., Hernando, C. E., Cuevas, J. C., Godoy Herz, M. A., Depetris-Chauvin, A., Simpson, C. G., Brown, J. W., Cerdan, P. D., Borevitz, J. O., Mas, P., Ceriani, M. F., Kornblihtt, A. R., and Yanovsky, M. J. (2010) A methyl transferase links the circadian clock to the regulation of alternative splicing, *Nature 468*, 112-116.

3. Stallcup, M. R. (2001) Role of protein methylation in chromatin remodeling and transcriptional regulation, *Oncogene 20*, 3014-3020.

4. Bedford, M. T. (2007) Arginine methylation at a glance, *J. Cell Sci. 120*, 4243-4246.

5. Lake, A. N., and Bedford, M. T. (2007) Protein methylation and DNA repair, *Mutat. Res. 618*, 91-101.

6. Smith, B. C., and Denu, J. M. (2009) Chemical mechanisms of histone lysine and arginine modifications, *Biochim. Biophys. Acta. 1789*, 45-57.

7. Liu, C., Lu, F., Cui, X., and Cao, X. (2010) Histone methylation in higher plants, *Annu. Rev. Plant Biol. 61*, 395-420.

8. Deng, X., Gu, L., Liu, C., Lu, T., Lu, F., Lu, Z., Cui, P., Pei, Y., Wang, B., Hu, S., and Cao, X. (2010) Arginine methylation mediated by the Arabidopsis homolog of PRMT5 is essential for proper pre-mRNA splicing, *Proc. Natl. Acad. Sci. U. S. A. 107*, 19114-19119.

9. Wolf, S. S. (2009) The protein arginine methyltransferase family: an update about function, new perspectives and the physiological role in humans, *Cell. Mol. Life Sci. 66*, 2109-2121.

10. Teyssier, C., Le Romancer, M., Sentis, S., Jalaguier, S., Corbo, L., and Cavailles, V. (2009) Protein arginine methylation in estrogen signaling and estrogen-related cancers, *Trends Endocrinol. Metab. 21*, 181-189.

11. Aletta, J. M., and Hu, J. C. (2008) Protein arginine methylation in health and disease, *Biotechnol. Annu. Rev. 14*, 203-224.

12. Boger, R. H., Cooke, J. P., and Vallance, P. (2005) ADMA: an emerging cardiovascular risk factor, *Vasc. Med. 10*, S1-2.

13. Ueda, S., Yamagishi, S., Matsumoto, Y., Fukami, K., and Okuda, S. (2007) Asymmetric dimethylarginine (ADMA) is a novel emerging risk factor for cardiovascular disease and the development of renal injury in chronic kidney disease, *Clin. Exp. Nephrol. 11*, 115-121.

14. Pullamsetti, S., Kiss, L., Ghofrani, H. A., Voswinckel, R., Haredza, P., Klepetko, W., Aigner, C., Fink, L., Muyal, J. P., Weissmann, N., Grimminger, F., Seeger, W., and Schermuly, R. T. (2005) Increased levels and reduced catabolism of asymmetric and symmetric dimethylarginines in pulmonary hypertension, *FASEB J. 19*, 1175-1177.

15. Bedford, M. T., and Richard, S. (2005) Arginine methylation an emerging regulator of protein function, *Mol. Cell. 18*, 263-272.

16. Lin, C. H., Hsieh, M., Li, Y. C., Li, S. Y., Pearson, D. L., Pollard, K. M., and Li, C. (2000) Protein N-arginine methylation in subcellular fractions of lymphoblastoid cells, *J. Biochem. 128*, 493-498.

17. Frankel, A., and Clarke, S. (2000) PRMT3 is a distinct member of the protein arginine N-methyltransferase family. Conferral of substrate specificity by a zinc-finger domain, *J. Biol. Chem. 275*, 32974-32982.

18. Cheng, X., Collins, R. E., and Zhang, X. (2005) Structural and sequence motifs of protein (histone) methylation enzymes, *Annu. Rev. Biophys. Biomol. Struct. 34*, 267-294.

19. Zhang, X., Zhou, L., and Cheng, X. (2000) Crystal structure of the conserved core of protein arginine methyltransferase PRMT3, *EMBO J. 19*, 3509-3519.

20. Zhang, X., and Cheng, X. (2003) Structure of the predominant protein arginine methyltransferase PRMT1 and analysis of its binding to substrate peptides, *Structure 11*, 509-520.

21. Kim, S., Merrill, B. M., Rajpurohit, R., Kumar, A., Stone, K. L., Papov, V. V., Schneiders, J. M., Szer, W., Wilson, S. H., Paik, W. K., and Williams, K. R. (1997) Identification of N(G)-methylarginine residues in human heterogeneous RNP protein A1: Phe/Gly-Gly-Gly-Arg-Gly-Gly-Gly/Phe is a preferred recognition motif, *Biochemistry 36*, 5185-5192.

22. Wooderchak, W. L., Zang, T., Zhou, Z. S., Acuna, M., Tahara, S. M., and Hevel, J. M. (2008) Substrate profiling of PRMT1 reveals amino acid sequences that extend beyond the "RGG" paradigm, *Biochemistry 47*, 9456-9466.

23. Gary, J. D., and Clarke, S. (1998) RNA and protein interactions modulated by protein arginine methylation, *Prog. Nucleic Acid Res. Mol. Biol. 61*, 65-131.

24. Osborne, T. C., Obianyo, O., Zhang, X., Cheng, X., and Thompson, P. R. (2007) Protein arginine methyltransferase 1: positively charged residues in substrate peptides distal to the site of methylation are important for substrate binding and catalysis, *Biochemistry 46*, 13370-13381.

25. Lin, W. J., Gary, J. D., Yang, M. C., Clarke, S., and Herschman, H. R. (1996) The mammalian immediate-early TIS21 protein and the leukemia-associated BTG1 protein interact with a protein-arginine N-methyltransferase, *J. Biol. Chem. 271*, 15034-15044.

26. Singh, V., Miranda, T. B., Jiang, W., Frankel, A., Roemer, M. E., Robb, V. A., Gutmann, D. H., Herschman, H. R., Clarke, S., and Newsham, I. F. (2004) DAL-1/4.1B tumor suppressor interacts with protein arginine N-methyltransferase 3 (PRMT3) and inhibits its ability to methylate substrates in vitro and in vivo, *Oncogene 23*, 7761-7771.

27. Pal, S., Vishwanath, S. N., Erdjument-Bromage, H., Tempst, P., and Sif, S. (2004) Human SWI/SNF-associated PRMT5 methylates histone H3 arginine 8 and negatively regulates expression of ST7 and NM23 tumor suppressor genes, *Mol. Cell. Biol. 24*, 9630-9645.

28. Xu, W., Cho, H., Kadam, S., Banayo, E. M., Anderson, S., Yates, J. R., 3rd, Emerson, B. M., and Evans, R. M. (2004) A methylation-mediator complex in hormone signaling, *Genes Dev. 18*, 144-156.

29. Niu, L., Lu, F., Pei, Y., Liu, C., and Cao, X. (2007) Regulation of flowering time by the protein arginine methyltransferase AtPRMT10, *EMBO Rep. 8*, 1190-1195.

30. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution, *Acta. Crystallogr. D Biol. Crystallogr. 66*, 213-221.

31. Fauman, E. B., Blumenthal, R. M., and Cheng, X. (1999) *Structure and evolution of Adomet-dependent methyltransferase. In S-Adenosylmethionine-Depedent Methyltransferase: Structures and Functions*.

32. Cheng, X., and Roberts, R. J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping, *Nucleic Acids Res. 29*, 3784-3795.

33. Lee, D. Y., Ianculescu, I., Purcell, D., Zhang, X., Cheng, X., and Stallcup, M. R. (2007) Surface-scanning mutational analysis of protein arginine methyltransferase 1: roles of specific amino acids in methyltransferase substrate specificity, oligomerization, and coactivator function, *Mol. Endocrinol. 21*, 1381-1393.

34. Smith, J. J., Rucknagel, K. P., Schierhorn, A., Tang, J., Nemeth, A., Linder, M., Herschman, H. R., and Wahle, E. (1999) Unusual sites of arginine methylation in Poly(A)-binding protein II and in vitro methylation by protein arginine methyltransferases PRMT1 and PRMT3, *J Biol Chem 274*, 13229-13234.

35. Teotico, D. G., Frazier, M. L., Ding, F., Dokholyan, N. V., Temple, B. R., and Redinbo, M. R. (2008) Active nuclear receptors exhibit highly correlated AF-2 domain motions, *PLoS Comput. Biol. 4*, e1000111.

36. Goulet, I., Gauvin, G., Boisvenue, S., and Cote, J. (2007) Alternative splicing yields protein arginine methyltransferase 1 isoforms with distinct activity, substrate specificity, and subcellular localization, *J Biol Chem 282*, 33009-33021.

37. Tang, J., Gary, J. D., Clarke, S., and Herschman, H. R. (1998) PRMT 3, a type I protein arginine N-methyltransferase that differs from PRMT1 in its oligomerization, subcellular localization, substrate specificity, and regulation, *J Biol Chem 273*, 16935-16945.

38. Weiss, V. H., McBride, A. E., Soriano, M. A., Filman, D. J., Silver, P. A., and Hogle, J. M. (2000) The structure and oligomerization of the yeast arginine methyltransferase, Hmt1, *Nat Struct Biol 7*, 1165-1171.

39. Pawlak, M. R., Banik-Maiti, S., Pietenpol, J. A., and Ruley, H. E. (2002) Protein arginine methyltransferase I: substrate specificity and role in hnRNP assembly, *J. Cell. Biochem. 87*, 394-407.

40. Stols, L., Gu, M., Dieckman, L., Raffen, R., Collart, F. R., and Donnelly, M. I. (2002) A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site, *Protein Expr Purif 25*, 8-15.

41. Cheng, Y., McNamara, D. E., Miley, M. J., Nash, R. P., and Redinbo, M. R. (2011) Functional characterization of the multi-domain F plasmid TraI relaxase-helicase, *J. Biol. Chem. 286*, 12670-12682.

42. Otwinowski, Z., and Minor, W. (1997) Processing of X-ray Diffraction Data Collected in Oscillation Mode, In *Methods in Enzymology* (Cater, C. W., Jr, and Sweet, R. M., Eds.), pp 307-326, Academic Press, New York.

43. Collabroative Computational Project Number 4. (1994) The CCP4 suite: programs for protein crystallography, *Acta Crystallogr. D Biol. Crystallogr. 50*, 760-763.

44. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. Features and development of Coot, *Acta. Crystallogr. D Biol. Crystallogr. 66*, 486-501.

45. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, *J. Comput. Chem. 24*, 1999-2012.

46. Case, D. A., Darden, T. A., Cheatham, T. E. I., Simmerling, C. L., and Wang, J. (2006) *AMBER 9*, University of California: San Francisco.

47. Essman, U., Perera, L., Berkowita, M. L., Darden, T. A., and Lee, H. (1995) A smooth particle mesh Ewald method. , *J Chem. Phys. 103*, 8577-8593.

48. Sharma, S., Ding, F., and Dokholyan, N. V. (2007) Multiscale modeling of nucleosome dynamics, *Biophys. J. 92*, 1457-1470.

49. Everitt, B. S., Landau, S., and Leese, M. (2001) *Cluster analysis*, Oxford University Press: Oxford.

CHAPTER 3

Crystal Structure of the HEAT Domain from the Pre-mRNA Processing Factor Symplekin

3.1 INTRODUCTION

Maturation of most eukaryotic pre-mRNAs requires cleavage and polyadenylation of the 3'-ends of primary transcripts. The 3'-end polyA tail ensures proper translation by delivering ribosomes to the mRNA(*1*); in amphibian oocytes, it was shown that translation was eliminated when the polyA tail addition was blocked by chemical modification(*2*). The polyA tail is also essential for protecting the message from exonucleases and for transporting the message from the nucleus to the cytoplasm(*3*). The length of the polyA tail affects the stability of the message, and compromised stability has been shown to lead to inflammation, cancer, early developmental maladies and coronary ailments(*4*). Thus, proper polyA tail addition to messenger RNA is required for proper cellular function.

For polyadenylation to occur, the cleavage stimulation factor (CstF) and the cleavage and polyadenylation specificity factor (CPSF) must work in concert to recognize and orient the cleavage site for the addition of the poly (A) tail(*5*). The ~1,160 residue Symplekin

Reprinted from Journal of Molecular Biology 392(1), Sarah A. Kennedy, Monica L. Frazier, Mindy Steiniger, Ann M. Mast, William F. Marzluff, and Matthew R. Redinbo, Crystal Structure of the HEAT Domain from the Pre-mRNA Processing Factor Symplekin, 115-28, 2009, with permission from Elsevier.

Structural coordinates have been deposited with the RCSB Protein Data Bank as 3GS3.

Monica Frazier contributed Section 3.3.5, a portion of Section 3.4, Table 3.2, and Figure 3.7.

protein is proposed to be the scaffolding factor on which this large protein complex is assembled(*3*). Symplekin binds two members of the CstF macromolecular complex, CstF64 and CstF77, in a mutually exclusive manner(*6*). Symplekin was identified as a stoichiometric component of the polyadenylation complex recently isolated from mammalian cells(*7*). Symplekin, CPSF73, and CPSF100 are part of a stable complex in *D. melanogaster* as shown via co-immunoprecipitation and co-depletion studies(*8*).

Metazoan replication-dependent histone mRNAs are unique in that their 3' ends are cleaved, but not polyadenylated. Interestingly, fractionation of HeLa cell nuclear extracts also identified Symplekin as a component of the histone pre-mRNA processing machinery(*9*). Additionally, an extensive RNA interference (RNAi) screen found Symplekin to be necessary for histone pre-mRNA processing in *D. melanogaster*; when Symplekin was RNAi-depleted, a histone pre-mRNA reporter(*10*) and endogenous histone mRNA(*8*) was misprocessed. These data lead to the hypothesis that Symplekin is essential for proper 3'-end formation of canonical and histone mRNA by providing a scaffold on which protein-protein interactions can occur (*6, 9*).

Symplekin may also serve as bridging factor between the polyadenylation machinery and transcription regulators. Most recently, the N-terminal region of yeast Symplekin (Pta1) was found to interact with Ssu72, an RNA polymerase II C-terminal domain (CTD) serine 5-phosphatase(*11*). The 124 N-terminal residues of mouse Symplekin interacts with heat shock factor 1 (HSF1). HSF1, Symplekin and other polyadenylation factors coimmunoprecipitate with HSF1 after heat shock, leading to the suggestion that HSF1 stimulates both transcription and processing (*12*). Over expression of a non-DNA binding mutant HSF1 to interfere with the HSF1-Symplekin interaction decreased Hsp70 mRNA polyadenylation in stressed cells (*12*). Thus, the N-terminal region of Symplekin may be involved in protein-protein interactions that help couple transcription and processing.

Utilizing *in silico* methods (*13-19*), several potential HEAT repeats were identified in the N-terminus of *D. melanogaster* Symplekin. Protein domains formed by HEAT repeats are established protein-protein interaction scaffolds(*20-27*). HEAT repeats are composed of 37-47 residues that fold into two anti-parallel helices connected by short (1-10 amino acids) linkers. Each set of helices can repeat 3 to 36 times, creating a HEAT domain(*16*). To characterize the N-terminal region of the Symplekins, the three-dimensional structure of *D. melanogaster* Symplekin residues 19-271 was determined using SAD phasing and refined to 2.4 Å resolution. Additionally, molecular dynamics simulations were employed to examine motion within this molecular scaffold. Taken together, these results provide the first detailed structural information on Symplekin, and indicate that the Symplekin HEAT domain may serve as a scaffold for protein-protein interactions essential to the mRNA maturation process.

3.2 RESULTS

3.2.1 Structure of the Symplekin HEAT Domain

Examination of the 1,165 residue *D. melanogaster* Symplekin sequence using secondary structure prediction algorithms indicated that a series of HEAT repeats are present in the first 300 amino acids of the protein, and that this domain was expected to be conserved in symplekin orthologues(*13-16, 19, 28*). The predicted *D. melanogaster* Symplekin HEAT domain (residues 19-271) was cloned and expressed in *E. coli*, purified to homogeneity and crystallized using hanging-drop vapor diffusion. The structure of the selenomethionine-substituted Symplekin HEAT domain was determined using SAD phasing methods to 2.9 Å resolution, and the structure of the native Symplekin HEAT domain was then refined to 2.4 Å resolution (Table 3.1). Figure 3.1a illustrates a portion of the Symplekin HEAT domain final model in the original 2.9 Å resolution experimental density from SAD

phasing. Residues 19-271 of *D. melanogaster* Symplekin contain five HEAT repeats that fold into a single domain with a crescent shape (Figure 3.1b). The ten HEAT helices (residues 22-256) are lettered conventionally for HEAT repeat domains (A for the convex and B for the concave surfaces). Repeats 1-5 contain 37, 37, 47, 46, and 42 amino acids, respectively, values similar to those observed for established HEAT repeats(*29*).

An extended 31-residue loop (amino acids 187-217 and denoted loop 8) connects helices 4B and 5A in the Symplekin HEAT domain structure. Six polar interactions are formed between this loop and helices 4B and 5A, as well as two internal hydrogen bonds that occur between residues within the loop (Figure 3.1c). Specifically, within the loop, a 2.8 Å hydrogen bond is formed between the backbone nitrogen of D192 and the side-chain oxygen of S195, and a 2.9 Å hydrogen bond is observed between the S203 backbone nitrogen and a D206 side-chain oxygen. Between the loop and the canonical HEAT domain scaffold, hydrogen bonds are observed between R258 of loop 10, M257 of α5B, K132 of α3B, and residues S195, G200, D201, and S203 of loop 8. Figure 3.2 illustrates the electrostatic potential of the concave surface of the molecule, indicating the presence of a positively charged patch as well as the predominantly negatively charged loop 8. The average thermal displacement parameter (B-factor) for loop 8 is 69 $Å^2$, while the overall average B-factor for the structure is 52 $Å^2$. One crystal contact involving loop 8 exists in the refined crystal structure, between D209 in loop 8 and E69 of α2A in a symmetry-related monomer.

3.2.2 Conservation in Symplekin Orthologues

In addition to the reported similarity between amino acids 300-800 of human and yeast Symplekin(*6*), the HEAT repeats within the N-terminal regions of Symplekin orthologues, including the residues on the concave surface and loop 8, are reasonably well conserved. Figure 3.3 presents a sequence alignment of the N-terminal ~300 residues of

Symplekins from eight representative species: *Drosophila melanogaster, Homo sapiens, Xenopus laevis, Strongylocentrotus purpuratus, Caenorhabditis elegans, Arabidopsis thaliana, Schizosaccharomyces pombe,* and *Saccharomyces cerevisiae.*(*30*). While only six amino acid positions (99, 152, 179, 180, 251 and 258) are 100% identical within the domain, 38% are highly similar (defined as 6 or more species containing a similar amino acid type). Of these similar residues, 75% are nonpolar and map to positions in the hydrophobic core of the *D. melanogaster* HEAT domain structure.

When comparing only the three sequences most closely related to *D. melanogaster* (*H. sapiens, X. laevis* and *S. purpuratus*), it was found that the hydrophobic core of Symplekin contains 28 residues that are completely conserved (Figure 3.4a; see also Figure 3.3). In addition, in considering the concave, convex and loop regions of Symplekin, it is evident that the majority of identical residues fall on the concave surface and within the loops (Figures 3.4b, 3.4c). The sixteen conserved residues found on the concave surface account for >20% of the total conserved residues in this HEAT domain (Figure 3.4b, yellow). Loop regions projecting from the concave surface account for ten conserved residues, five of which are in loop 8 (Figure 3.4b, cyan). In contrast to the concave side, the convex surface contains only 4 conserved residues (Figure 3.4c, green). These data indicate that the HEAT domain is likely conserved in the N-terminal regions of the Symplekins of known sequence, and that the hydrophobic core, concave surface and loop 8 are the regions most highly conserved.

Although sequence variation exists at many positions in the more distant species (sequences in grey, Figure 3.3), examination of secondary structure predictions indicates that the helical HEAT-like fold is preserved in each putative Symplekin orthologue(*31*). All seven sequences have unstructured regions aligning with *D. melanogaster* loops, including the extended loop 8 (underlined in Figure 3.3). While *S. cerevisiae* secondary structure predictions in the regions of α2B and α4A include sequence inserts, homology modeling

supports the conclusion that this protein adopts a HEAT-like repeat structure. Taken together, these data indicate that the orthologue sequences shown in Figure 3.3 are likely to resemble the α-helical *D. melanogaster* Symplekin HEAT domain structure.

3.2.3 Symplekin HEAT Repeats are Classified with Scaffolding Proteins

The closely related HEAT and armadillo structural domains have been sub-classified based on specific amino acid sequences that coincide with functional categorization(*20*). To further characterize the Symplekin's N-terminal domain, each of the repeats were structurally aligned and the sequences were compared to the sequence classifications for three types of HEAT sequences (ADB, AAA and IMB), as developed by Andrade *et al.*(*20*). The AAA, ADB, and IMB HEAT classes all exhibit a similar pattern of hydrophobic residues and contain conserved residues D19 and R/K 25 near the intrahelical loop, while the sequence logo of the ADB class also contains D/N21 and V/I24(*20*). Symplekin contains the ADB pattern: HEAT repeat 2 contains D77, N79, V92, and K83, HEAT repeat 3 includes D114, N115, I120, and K121, while HEAT repeat 4 contains 167D, 170N, 173I and R174. Terminal HEAT repeats are more difficult to classify because they have a different set of packing constraints(*20*). The highly conserved P11 of the AAA and IMB classes is lacking in the ADB class, and is also lacking in the HEAT repeats 2, 3 and 4 of Symplekin. Taken together, the residues in the three central Symplekin HEAT repeats indicate that Symplekin may belong to a small ADB subclass of HEAT repeats, a family containing mainly α, β-adaptin and β-coat proteins that function as scaffolds for protein binding and transport. This sub-classification supports the hypothesis that the Symplekin HEAT domain has a structure appropriate for protein-protein interactions.

3.2.4 Symplekin HEAT Structurally Aligns with Protein-Binding Scaffolds

The structure of the *D. melanogaster* Symplekin HEAT Domain was examined using Dali to identify proteins of similar structure(*32*). While nearly 200 protein structures exhibited homology with the Symplekin HEAT Domain, the closest structural neighbors were serine/threonine-protein phosphatase 2A PR65/A subunit (PDB 1b3u), Cullin-associated protein Cand1 (PDB 1u6gc), and karyopherin-α (PDB 1ee4), all of which have HEAT or armadillo (ARM) repeats. Experimental evidence indicates that their HEAT/ARM repeats are involved in protein-protein interactions and the majority of these domains utilize their concave face as a protein binding or scaffolding surface(*22, 24, 27, 33-36*). Recall that amino acid conservation supported the functional importance of the concave surface and loop 8 of the Symplekin HEAT domain (see above). Symplekin superimposes on the structure of Cand1 (TIP120) of the Cand1-Cul1 complex with only 10% sequence identity but with 3.8 Å RMSD over 203 aligned residues, and a Z-score of 14.7 (Figure 3.5a). The concave surface of Cand1 is employed in binding Cul1 to inhibit Cul1 from forming the E3 ubiquitin ligase complex(*22*). Symplekin structurally superimposes on yeast karyopherin-α with 11% identity over 196 Cα positions, a 5.0 Å RMSD, and a Z-score of 14.2 (Figure 3.5b)(*32*). The concave surface angles of each protein were calculated by measuring the angle between three concave surface Cα residues on helices 1B, 3B and 5B at three positions on these helices: near the N-terminus, the center, and near the C-terminus. The concave surface angle for the helical N-termini of Symplekin, Cand1 and karyopherin-α are 144°, 100°, and 153°, respectively; for the helical centers are 141°, 124°, and 157°, respectively; and for the helical C-termini are 107°, 137°, and 150°, respectively. The twist of each HEAT domain was determined by comparing the angle between the helical axes of helices 1B and 5B, and found to be 5°, 10°, and 77° for Symplekin, Cand1 and karyopherin-α, respectively. Thus, while the overall concave surface angles of each HEAT domain are

similar, Symplekin and Cand1 exhibit significantly less domain twist than does karyopherin-α.

The core of yeast karyopherin-α is a canonical ARM repeat with acidic concave surface regions equipped to bind basic nuclear localization signals (NLS)(*36*). It has been reported that *D. melanogaster* karyopherin-α3 binds to the positively charged NLS of HSF1, and residues 1-124 of Symplekin interact with HSF1(*12, 37*). Fly and yeast karyopherin-α sequences share 50% identity and maintain a similar electrostatic surface. There are no extended loops in either karyopherin-α sequence. However, with respect to loop 8, it is clear from the structural superposition of karyopherin-α and the Symplekin HEAT domain that the position of loop 8 clashes with the NLS sequence bound to the surface of karyopherin-α (Figure 3.5b). Loop 8 of Symplekin is negatively charged and could provide an alternative binding region for the positively charged NLS (Figures 3.2, 3.5b). Taken together, the observations that karyopherin-α3 and Symplekin contain similar structural motifs, have similar electrostatic surfaces, and both bind to HSF1 support the conclusion that Symplekin has characteristics of a protein-binding scaffold.

3.2.5 Loop 8 Impacts Symplekin HEAT Domain Motion

Molecular dynamics (MD) simulations have been used to investigate the manner in which HEAT and ARM domains change conformational states upon ligand binding, and to design ideal ARM domains for general peptide binding(*38-40*). Our attempts at biochemically characterizing the interactions between the Symplekin HEAT domain with *D. melanogaster* CstF64 and Ssu72 through amylose-affinity pull down assays were unsuccessful due to non-specific interaction with the MBP tag. However, these interactions have been shown indirectly in Symplekin orthologues. Instead, we employed MD to examine how loop 8 impacts the overall and correlated motions within the Symplekin HEAT domain

structure. Three models of the Symplekin HEAT domain were subjected to 10 ns MD simulations: Wild-Type containing a complete loop 8, a model in which the ten polar residues in loop 8 were all replaced with serine (Poly-Ser Loop 8; serine was chosen to place small polar side chains in this surface-exposed loop), and a model in which loop 8 is replaced with a short turn (Short Loop 8) (Figure 3.6). The Short Loop 8 mutant was designed with the intention to mimic the minimal loops commonly seen between HEAT repeats. Comparing the Short Loop 8 with Wild-Type Symplekin was expected to show the role loop 8 plays in the motion of the Symplekin HEAT domain. The Poly-Ser Loop 8 model was expected to show whether specific residues on the loop were important for Symplekin HEAT domain motion.

Simulations of each Symplekin model were performed in triplicate using different random number generator seeds. Data used for analysis of each individual simulation was collected from 10 consecutive nanoseconds of the same conformational ensemble (designated by a consistent root mean square deviation from the starting crystal structure) (Figure 3.7a). The models were analyzed with respect to both the overall degree of motion seen in Cα atoms (observed as the atomic position fluctuations (APF) of each Cα) as well as the behavior of each Cα with respect to all other Cα atoms. Wild-Type loop 8 and Poly-Ser loop 8 simulations exhibit nearly identical overall motion in loop 8 Cα atoms as well as throughout the entire protein (Table 3.2). The similarity of the mean APFs between the Wild-Type loop 8 and the Poly-Ser loop 8 indicates that the specific amino acids in loop 8 do not control the overall motion in either loop 8 or the entire HEAT domain. The Short Loop 8 simulation's overall degree of motion was also found to be similar to both the Wild-Type loop 8 and Poly-Ser loop 8 simulations, indicating that the presence of the extended loop 8 does not significantly influence the overall motion of the HEAT domain.

Correlation-anticorrelation plots, which provide information on the relative motion of each residue pair during an MD trajectory, were then generated for these HEAT domain simulations. In Figures 3.7b-d, red indicates correlated motion between two Cα positions (*e.g.,* motion in the same direction), blue indicates anti-correlated motion (*e.g.,* in the opposite direction), and yellow indicates no correlation in motion (two residues that move randomly with respect to one another). Both the Wild-Type and Poly-Ser Symplekin HEAT domain simulations exhibit similar patterns and levels of correlated and anticorrelated motion (Figures 3.7b, 3.7d), indicating that the dynamics of the HEAT domain is maintained regardless of the specific residues present in loop 8. In contrast, however, the Short loop 8 simulation exhibits noticeably higher levels of correlated and, particularly, anticorrelated motions (Figure 3.7c), indicating that removal of the loop increases the degree of specific residue-to-residue motions within the HEAT domain. Taken together, these results indicate that the presence of loop 8, but not specific polar residues on the loop, reduces specific pairwise motions in the Symplekin HEAT domain. Thus, maintaining an extended loop in this location in Symplekin (*e.g.,* see Figure 3.3) may disrupt specific domain movements to provide the neutral scaffold for protein-protein interactions.

3.3 DISCUSSION

The 1165-residue Symplekin protein is a component of the 3'-end processing machinery critical to both canonical and histone messenger RNA(*3, 6, 9*). While structural information is available for many of the other 3'-end processing factors(*41-49*), no structures have been reported for any region of Symplekin to date. Here, we show that residues 19-271 of *D. melanogaster* Symplekin fold into a HEAT repeat structure with an extended loop 8 that is conserved in the Symplekins of known sequence. Examination of the electrostatic

potential of the Symplekin HEAT domain reveals that the concave surface is positively charged, while the ridge formed by the even-numbered loops exhibits a slight overall negative charge (Figure 3.2). Sub-classification of Symplekin's HEAT repeats and structural alignments indicate that these regions of Symplekin may act as a scaffold for protein-protein interactions. Indeed, HEAT domains are well established platforms for macromolecular complex formation (*e.g.,* Figure 3.5). For example, crystal structures and molecular dynamics studies of importin-β reveal four regions for peptide binding within 5 HEAT repeats(*50*).

Takagaki *et al.* has reported that the central region (residues 300-740) of human Symplekin is 31% similar to *S. cerevisiae* Symplekin orthologue, Pta1(*6*). Using the crystal structure reported here as a guide, we further examined Symplekin orthologue sequences and have found that the N-terminal region of Pta1 exhibits some homology to the equivalent region of *D. melanogaster* Symplekin. All of the orthologues contain similar hydrophobic/hydrophilic residue distributions and have greater than 60% α-helical content in their N-terminal regions (Figure 3.3). Specifically, Pta1 maintains hydrophobic residues in 79 out of the 125 hydrophobic positions present in the *D. melanogaster* N-terminal HEAT domain and 33 residues are identical between these two species. Loop 8 lacks secondary structure in all species investigated, and three residues are identical and nine residues are similar between Pta1 and *D. melanogaster* Symplekin within this 31-residue region. These data support the conclusion that the N-terminus of *S. cerevisiae* Pta1 likely encodes α-helical HEAT-like domain similar to the *D. melanogaster* HEAT domain structure reported here.

Several published reports map specific protein docking sites within the Symplekin HEAT region. The HEAT domain of the yeast Symplekin homologue Pta1 has been shown to bind to both Ssu72 and Glc7(*11, 51*) and a portion of the mouse Symplekin HEAT domain interacts with HSF1(*12*). Ssu72 and Glc7 have been implicated in the regulation of 3'-end

processing. Depletion of the Glc7 phosphatase causes an accumulation of phosphorylated Pta1 and a subsequent reduction in 3'-end polyadenylation; this effect can be rescued by the addition of either Glc7 or unphosphorylated Pta1 back into the processing reaction(*51*). The binding of yeast Symplekin homologue Pta1 to Ssu72, an RNA polymerase II C-terminal domain phosphatase, may position the 3'-end processing machinery in proximity to primary transcripts to promote facile processing(*11*). Similarly, Symplekin may link 3'-end processing to transcriptional control via contacts with transcription factors like HSF1. The binding of the HEAT domain of mouse symplekin to HSF1 promotes polyadenylation of Hsp70 mRNA in heat stressed cells(*11, 12*). Taken together, these data indicate that the Symplekin HEAT region provides a platform for enzymes and other proteins critical to modulating 3'-end processing.

GST pull down studies and yeast two-hybrid assays provide information on the specific Symplekin regions involved in these protein-protein interactions (Figure 3.8)**.** Binding to Glc7 was maintained using Pta1 Δ1-100, while removal of Symplekin residues 1-200 abolished Glc7 binding(*51*). Indirectly, this indicates Symplekin HEAT repeats 3, 4 and loop 8 (Pta1 residues 100-200) are used in binding to Glc7(*51*). Ssu72 requires Symplekin HEAT repeat 2 for optimal binding (Pta1 residues 51-76)(*11*), and HSF1 binds to residues HEAT repeats 1-3 (mouse Symplekin 1-124)(*11, 12, 51*). The exact regions of Symplekin required for interacting with the core 3'-end processing machinery, CstF and CPSF, have not been determined; however, it has been shown that some processing in yeast can occur with a Δ1-300 Pta1 construct(*11*). Therefore, we propose a model where several regulatory proteins bind in a mutually exclusive manner to distinct sites on the Symplekin HEAT domain, whereas the C-terminal region of the protein associates with central members of the 3'-end processing machinery (Figure 3.8).

Molecular dynamics simulations conducted on the *Drosophila* Symplekin domain structure provides preliminary insight into the motions in this region of the protein. Although

the timescale on the trajectories were only 10 nsec and the domain was examined in isolation, it was clear that the loop 8 is involved in disrupting the dynamic relationship between the residues across the entire HEAT domain (Figure 3.7). These observations suggest that the wild-type Symplekin HEAT domain may be tuned to adopt a more neutral range of motions to prepare it for binding to different protein partners. Changes in flexibility of wild-type proteins relative to specific mutants have been reported in previous molecular dynamics studies(*52*). Additionally, our characterization of the Symplekin HEAT domain agrees well with published MD studies of Armadillo and HEAT domain proteins. Examination of Cse1p by MD indicates that a particularly negatively charged loop (insert 19) helps to poise the structure in an open conformation to facilitate binding to RanGTP and Kap60p(*39*). Loop 8 in *D. melanogaster* Symplekin also exhibits a slight overall negative charge and may play a similar role in preparing the domain to bind to protein partners Glc7, Ssu72 or HSF1. In simulations of importin-β, the ligand bound states are curved in shape, but upon ligand release the domain opens to produce more elongated states(*39, 40*). The Symplekin HEAT domain may also employ such "tertiary disorder"(*40*) in conforming to different protein-binding partners. It is likely that there may be additional partners that interact with the HEAT domain that may be involved in regulating histone pre-mRNA processing.

Combining our structural and molecular dynamics results with biochemical studies, we have classified the Symplekin HEAT domain as a scaffold for the binding of proteins critical to modulating 3'-end mRNA processing. Utilizing sequence conservation data (Figures 3.3, 3.4), future biochemical and mutagenesis studies will be conducted with this HEAT domain to identify specific residues vital for binding to Glc7, Ssu72 and HSF1. A preliminary cryo-EM image of the purified 3'-end processing complex including Symplekin, CPSF, CstF and CFI has recently been determined at low resolution(*7*) and crystal structures exist for several components of the eukaryotic 3'-end processing machinery,

including CPSFs 30, 73, 100, CstF64 and 77 and CFI_m-25(*41, 42, 46, 47*). Thus, a range of efforts are underway to understand the intricate macromolecular relationships required for the catalytic and regulatory aspects of 3'-end processing machinery. The structure of the Symplekin HEAT domain presented here provides an additional piece of this complex structural puzzle.

3.4 METHODS

Expression and Purification of Symplekin HEAT Domain

The following software programs were utilized to predict the structural elements within Symplekin: BLAST, Jpred(*15*), PHYRE(*17*), pFam(*16*), InterProScan(*19*), ScanSite(*53*), PredictProtein(*18*), RONN(*54*) and COILS(*55*). The disordered regions include 1-18, 452-544, and 1116-1165. A HEAT-like domain was predicted between residues 19-271. Based on these analyses, residues 19-271 of *D. melanogaster* Symplekin were cloned into the expression vector pMCGS9, which provided N-terminal 6-histidine and maltose-binding protein (MBP) tags followed by a Tobacco Etch Virus (TEV) protease site(*56*). *Escherichia coli* BL21 (DE3) gold cells (Stratagene) were transformed with this constructed plasmid and cells were grown at 37 °C in 1.5 L of terrific broth supplemented with 50 mg/L ampicillin until an A_{600}=1.0-1.2. The temperature was dropped to 18 °C and 0.1 mM of IPTG was added to induce protein expression until a final OD A_{600}=4.5. The cells were harvested by centrifugation and resuspended in nickel buffer A (5 mM imidazole, 50 mM potassium phosphate, pH 7.4, 150 mM NaCl, 1 mM DTT, 0.01% sodium azide) and stored at -80°C. Thawed cells were lysed by sonication in the presence of DNase and protease inhibitors, and centrifuged at high speed for 60 minutes to produce a cleared lysate. The histidine-tagged protein was purified from the lysate by nickel affinity chromatography. Nickel buffer B (500 mM imidazole, 50 mM potassium phosphate, pH 7.4,

150 mM NaCl, 1 mM DTT, 0.01% sodium azide) was used to elute the protein from the column with a gradient of 5-100% B. To cleave the 6xHis-MBP fusion protein from the Symplekin 19-271 polypeptide, 2% TEV protease by mass TEV/mass Symplekin was added. Protein was dialyzed into nickel buffer A during TEV cleavage. A second nickel column purified the now un-tagged Symplekin from the 6×His-MBP tag. A polishing step of size exclusion chromatography (Column: Superdex 75, GE Healthcare; sizing buffer: 10 mM HEPES, pH 8.0, 50 mM NaCl, 1 mM DTT and 0.01% sodium azide) produced >95% purity by SDS PAGE. A selenomethionine-substituted form of *D. melanogaster* Symplekin residues 19-271 was produced using B834 cells, a methionine auxotroph cell line. Cells were grown in selenomethionine specific media (Athena) supplemented with 50 mg/L selenomethionine. Expression and purification procedures were identical to those listed above for the native protein.

Crystallization and X-ray Data Collection

Native and selenomethionine-substituted Symplekin proteins were concentrated to 3-6 mg/mL in sizing buffer. Crystallization was performed by hanging drop diffusion at 22 °C with mother liquor consisting of 0.4-0.5 M sodium citrate, 25-28% PEG 3350, 10 mM HEPES, pH 8.0, 0.01% N_3Na and 1 mM DTT. Each crystallization drop contained 1 µL of protein and 1 µL of well solution. Diamond shaped crystals grew within one week, with maximal dimensions of 300 µm x 60 µm x 60 µm. Crystals were cryoprotected in mother liquor plus 35% PEG 3350 and flash-cooled in liquid nitrogen. Diffraction data were collected at 100K using Sector 22-BM (SER-CAT) of the Advanced Photon Source at Argonne National Laboratories. A SAD data set was collected on crystals containing selenomethionine-substituted protein at 0.97190 Å; a native data set was collected using crystals containing wild-type protein at 0.97958 Å. DENZO and SCALEPACK in HKL-2000

were employed for data indexing and scaling(*57*). The crystals were of the space group $P4_12_12$ with unit cell dimensions of a, b = 68.7 Å, c = 138.5 Å and α, β, γ = 90° (Table 3.1).

Phasing, Model Building and Refinement

The SGXPRO software package, an interface for programs including SHELXD and SOLVE/RESOLVE, was employed to identify heavy atom sites and provide initial phases(*58*). A Matthews's coefficient value of 2.9 indicated that 1 molecule was expected in the asymmetric unit with 57.6% solvent. Six methionine residues were present in Symplekin 19-271, thus six Se sites were expected. SHELXD and SOLVE identified all six Se atom positions, and initial phases were calculated to 2.9 Å. RESOLVE was used for density modification and to provide an initial model. After these steps, the overall figure of merit was 0.69.

The model was built further by hand using COOT(*59*). Initially, all helices were built with alanine residues. Loops were added over several rounds of refinement to connect the helices. Finally, side chains were placed in the model. This 2.9 Å model from SAD was refined using REFMAC5 at this stage to R and R_{free} values of 0.353 and 0.419, respectively. To phase the 2.4 Å native data set, the model refined using the SAD data was used in molecular replacement(*60*). Further refinement was conducted by building and validating the model in COOT, and employing both CNS and REFMAC5 to produce R and R_{free} values of 0.2068 and -0.2653, respectively (Table 3.1). For both the original SAD data and the final native data, 5% of the data were set aside for the free-R and not used at any stage of refinement. The final model, consisting of 248 residues (no density was present for residues 19-21 and 271) and 142 water molecules, was validated with PROCHECK and Molprobity(*61*). Figures 3.1, 3.2, 3.4, 3.5 and 3.6 were created using PyMOL(*62*).

Sequence and Structural Alignments

The amino acid sequence of D. melanogaster Symplekin was entered into NCBI BLAST to retrieve homologous protein sequences. Sequences (with NCBI Accession numbers) from *Drosophila melanogaster* (NP_649580.1), *Homo sapiens* (NP_004810.2), *Xenopus laevis* (NP_001079691.1), *Strongylocentrotus purpuratus* (XP_783721.2), *Caenorhabditis elegans* (NP_505210.2), *Arabidopsis thaliana* (NP_195760.1), *Schizosaccharomyces pombe* (NP_594351.2) and *Saccharomyces cerevisiae* (AAA34919.1) were selected to represent a broad spectrum of species containing Symplekin. The sequence alignment, prepared using ClustalX and refined using several rounds of PSI-BLAST, was abbreviated to display only the portion of the sequences that align with the *D. melanogaster* HEAT domain structure (Figure 3.3). The structural alignment shown in Figure 3.4 was prepared using Dali(*32*). To characterize the HEAT repeats, each Symplekin HEAT repeat was structurally aligned to HEAT repeat 2.

Molecular Dynamics Simulations

COOT(*59*) was utilized to create the Symplekin modeled Short loop 8 and Poly-Ser loop 8. For the Poly-Ser model, the residues changed to serine were D192, E193, D194, K197, R198, D199, D201, D209, H210, R215. To design a short turn to replace loop 8 in the Short loop model, many other HEAT repeat proteins were examined to identify common linkers and it was determined that six residues are sufficient to bridge a 10.6 Å gap. To keep this linker as authentic as possible, residues on each end of the loop were maintained and connected with a glycine, a residue common in loops of HEAT repeats. Native residues 190-214 were completely removed. Thus, the modeled Short loop 8 is 187-LQSGRR-216. Residues 187-216 were used to calculate the relative APF values for loop 8 in each simulation.

Molecular dynamics simulations of the Symplekin HEAT domains were performed in triplicate using the AMBER 2003 force field with at 2 fs time step(*63*). LEaP was used to generate the topology and parameter files, SANDER performed the 5000 steps of energy minimization, which included constant volume followed by constant temperature equilibration, the PMEMD module was used for the production runs, and PTRAJ was utilized for analysis of the results(*63*). TIP3P water molecules were used to generate the solvated structure(*64*), and electrostatic interactions were calculated using the particle-mesh Ewald algorithm with a cutoff of 10 Å applied to Lennard-Jones interactions(*65*). All molecular dynamics simulations were conducted and analyzed as described previously(*66*).

3.5 ACKNOWLEDGEMENTS

This study was financially supported by NIH R01 grant DK62229 (M.R.R.) and the Graduate Assistants in Areas of National Need (GAANN) fellowship (S.A.K.). Data were collected at Southeast Regional Collaborative Access Team (SER-CAT) beamline 22 at the Argonne National Laboratory Advanced Photon Source. Use of the Advanced Photon Source was supported by the Office of Basic Energy Sciences of the U.S. Department of Energy Office of Science under contract no. W-31-109-Eng-38. We thank E. Ortlund, R. Duronio, D. Tatomer, L. Charlton, J. Orans, B. Wallace, K. Brennaman and A. Tripathy for helpful discussions.

3.6 FIGURE LEGENDS

Figure 3.1 Symplekin HEAT domain structure. (a) A wall-eyed stereo view representation of a portion of the final model in the original experimental electron density from SAD phasing contoured to 1σ. **(b)** The overall structure of the HEAT domain within Symplekin. Helices are lettered and numbered according to classical HEAT naming; the A helices create the

convex face, while the B helices create the concave face. The rainbow denotes the N to C progression of residues 22-270. The B helices are in light colors corresponding to their counterpart A helices. For example, 1A is red and 1B is pink. The extended loop 8 is in cyan. **(c)** Polar contacts within the loop 8 region of Symplekin. Arginine 258 and aspartic acid 201 form a salt bridge that anchors loop 8 to helix 5B. Lysine 132 forms a salt bridge with G200 to hold the loop in place with respect to helix 3B. A variety of other polar contacts position the extended loop 8 at the ends of helices 3-5 including S195-D192, M257-S195, M257-R258, R258-S203, S203-D206.

Figure 3.2 Electrostatic representation of the concave surface of Symplekin's HEAT domain. Red denotes negatively charged surfaces, blue denotes positively charged surfaces. The molecule is rotated 90° along the horizontal axis of Figure 3.1b, to orient the concave surface towards the reader. The concave surface is mainly positively charged, while loop 8 is negatively charged.

Figure 3.3 Sequence alignment of Symplekin orthologues in various species. The secondary structure elements and numbering across the top of the sequences correspond to the *D. melanogaster* structure in Figure 3.1c. Pink blocks denote the conserved D/E19 and K/R25 required for HEAT repeats, and blue colored blocks represent the HEAT repeat hydrophobic signature. The more distantly related orthologue sequences are shown in grey. Sequences and alignment were made using PSI-Blast and ClustalX. Secondary structure prediction and models of each sequence were predicted using PHYRE. Black underline denotes regions of disorder predicted by PHYRE, non-underline sequences are all predicted to be α-helical.

Figure 3.4 Conserved residues among four closely related Symplekin orthologues (*H. sapiens, X. laevis, D. melanogaster, S. purpuratus*) mapped onto the HEAT domain structure. (a) View of the hydrophobic core where purple represents residues with 100% conservation. (Molecule is rotated 180° on the vertical axis with respect to Figure 3.1b.) **(b)** View of the concave surface colored as follows: yellow denotes 100% conserved residues that project out of the concave surface, cyan residues are conserved in loop regions, and gray residues are not 100% conserved. (Molecule is rotated 90° on the horizontal axis with respect to Figure 3.4a.) **(c)** View of the convex surface colored as in A, except green denotes conserved residues that project from the convex surface.

Figure 3.5 Symplekin structural alignment with two most closely related structures. (a) Symplekin superimposed with Cand1 of the Cand1-Cul1-Roc1 complex (PDB 1u6g). Cand1 structure is in grey, Cul1 in white, and Roc1 is removed for figure clarity. Symplekin helices and surface have coloring from Figure 3.1b. A closer look at aligned individual helices shows that α3B is extended in comparison to the aligned helix in Cand1. Loop 8 is unique to Symplekin compared to Cand1. **(b)** Symplekin superimposed with karyopherin-α (PDB 1ee4). Karyopherin-α is grey, the nuclear localization signal (NLS) peptide bound to karyopherin-α is magenta. Symplekin maintains coloring from Figure 3.1b. Loop 8 docks into α-helices 3B, 4B and 5B, and lies in the same region that the NLS peptide occupies on karyopherin-α.

Figure 3.6 Symplekin HEAT domain structures used for molecular dynamics simulations. The labeled residues in loop 8 are mutated to serine for the Poly-Ser Loop 8 simulation. To prepare the short loop model, residues 191-214 were removed and wild-type residue 189 was connected to 215 by mutating F190G.

Figure 3.7 Truncation of loop 8 increases correlation/anticorrelation within Symplekin's HEAT domain. (a) All atom root mean squared deviation in position over the simulation time scale. Boxed area between 5-15 ns represents section of time used in data analysis. **(b-d)** Correlation/anticorrelation plots for Wild-Type **(b)**, Short Loop 8 **(c)**, and Poly-Ser Loop 8 **(d)** from molecular dynamics simulations. Red represents correlated movement, blue represents anticorrelated movements and colors between represent less correlated movements according to the given color scale. Each axis represents the Cα position for the given residue within the HEAT domain (going from N- to C- terminus from both left to right and from bottom to top). The secondary structural elements are colored consistently with Figure 3.1b.

Figure 3.8 Symplekin model for protein scaffolding. A diagram illustrating Symplekin HEAT domain's interaction with known binding partners. Secondary structural elements and residue numbers are labeled according to the structure. HSF1, Ssu72 and Glc7 bind to specific regions of the HEAT domain as described in the text. The C-terminal region of Symplekin has yet to be structurally characterized with respect to binding to the core 3'-end machinery.

Table 3.1

Data collection, phasing, and refinement statistics.

Data collection		
X-ray source	APS SER-CAT BM-22	
Space Group	$P4_12_12$	
Unit cell a,b,c (Å); α, β, γ (°)	68.7, 68.7, 138.5; 90, 90, 90	
Data set	SeMet	Native
Wavelength (Å)	0.97190	0.97958
Resolution (Å) (highest shell)	50.0-2.9 (3.0-2.9)	50.0-2.4 (2.49-2.40)
R_{sym}	9.4 (34.4)	8.0 (41.9)
I/σ	22.4 (1.0)	24.8 (1.9)
Completeness (%)	78.1 (6.7)	96.1 (79.6)
Redundancy	10.4 (1.6)	6.4 (2.8)
Phasing		
Mean Figure of Merit		
Centric	0.71	
Acentric	0.68	
All	0.69	
Refinement		
Resolution (Å)	50.0-2.4	
No. reflections	12465	
R_{work}	0.2068	
R_{free}	0.2653	
Molecules per asymmetric unit (AU)	1	
No. of amino acids per AU	248	
No. of waters per AU	142	
Average B-factors	46.37	
R.M.S. deviations		
Bond lengths (Å)	0.0059	
Bond angles (°)	1.20	
Ramachandran (%)		
Favored	96.76	
Outliers	0.40	

•$R_{sym} = \Sigma|I-I_{mean}|/\Sigma I$ where I is the observed intensity and I_{mean} is the average intensity of several symmetry related observations.

•$R_{work} = \Sigma|F_o-F_c|/\Sigma F_o$ where F_o and F_c are the observed and calculated structure factors, respectively.

•R_{free}=calculated as above for 5% of data not used in any step of refinement.

Table 3.2

Atomic position fluctuations (Å^2) for all Cα or loop Cα atoms.

Model	Wild-type loop 8		Poly-Ser loop 8		Short loop 8
Residues	All	Loop 8	All	Loop 8	All
Mean	1.0 ± 0.024	1.2 ± 0.035	1.1 ± 0.087	1.4 ± 0.32	1.2 ± 0.036
Max	3.5 ± 0.57	1.90 ± 0.32	3.8 ± 1.1	1.6 ± 0.095	3.8 ± 0.80
Min	0.48 ± 0.014	0.52 ± 0.036	0.52 ± 0.020	0.54 ± 0.046	0.59 ± 0.031

Figure 3.1

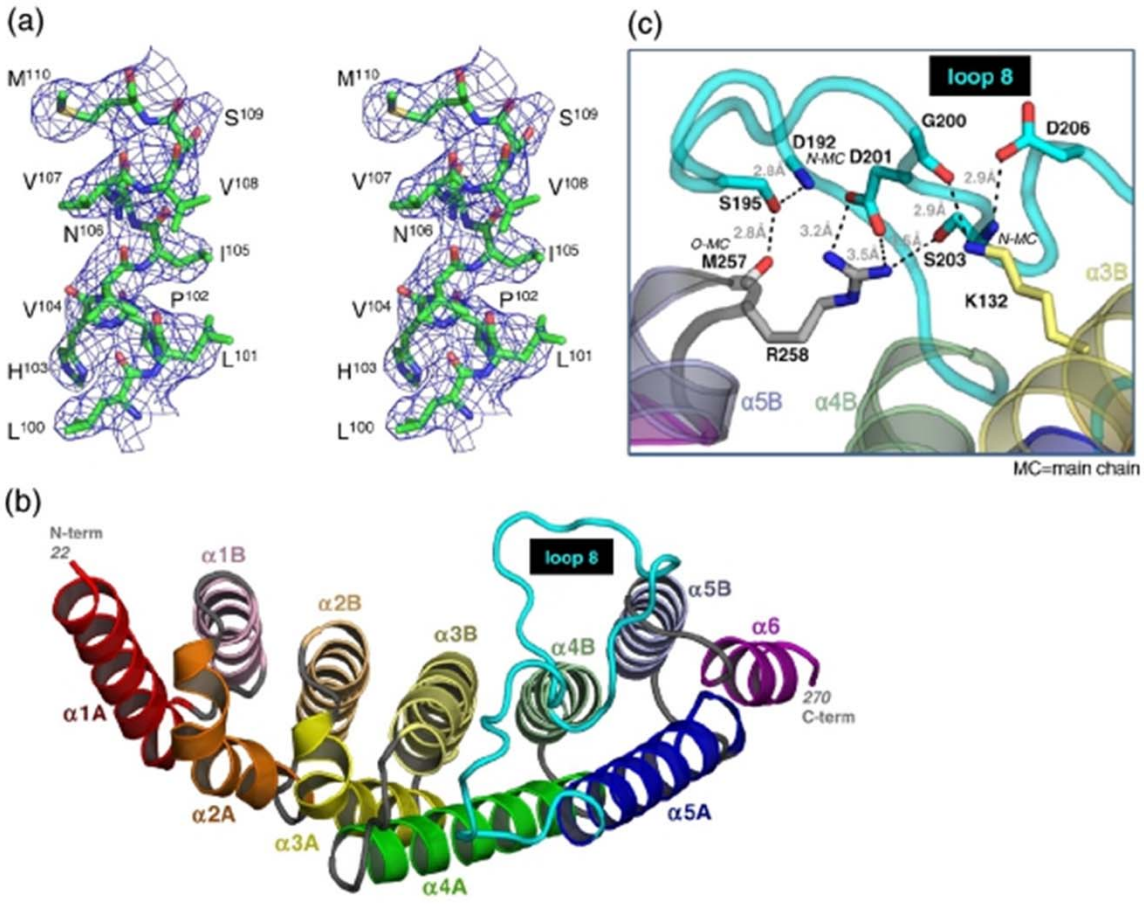

(a)
M110
S109
V107
V108
N106
I105
V104
P102
H103
L101
L100
(c)
loop 8
G200
D206
D192
N-MC D201
2.8Å
2.9Å
S195
2.9Å
O-MC
2.8Å
3.2Å
N-MC
M257
3.5Å
S203
α3B
K132
R258
α5B
α4B
MC=main chain
(b)
N-term
22
α1B
loop 8
α2B
α3B
α4B
α5B
α6
270
C-term
α1A
α2A
α3A
α4A
α5A

Figure 3.2

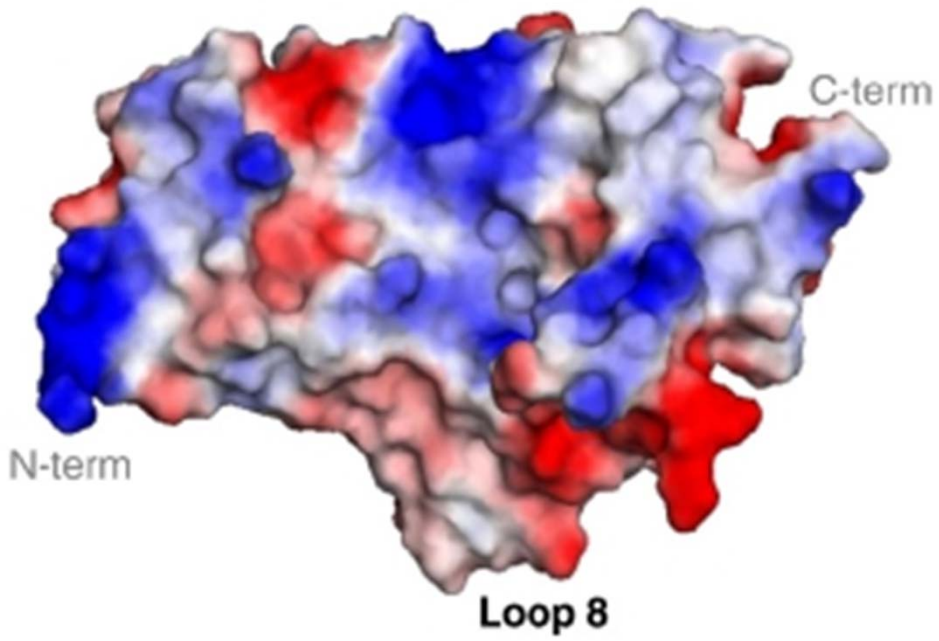


Figure 3.3

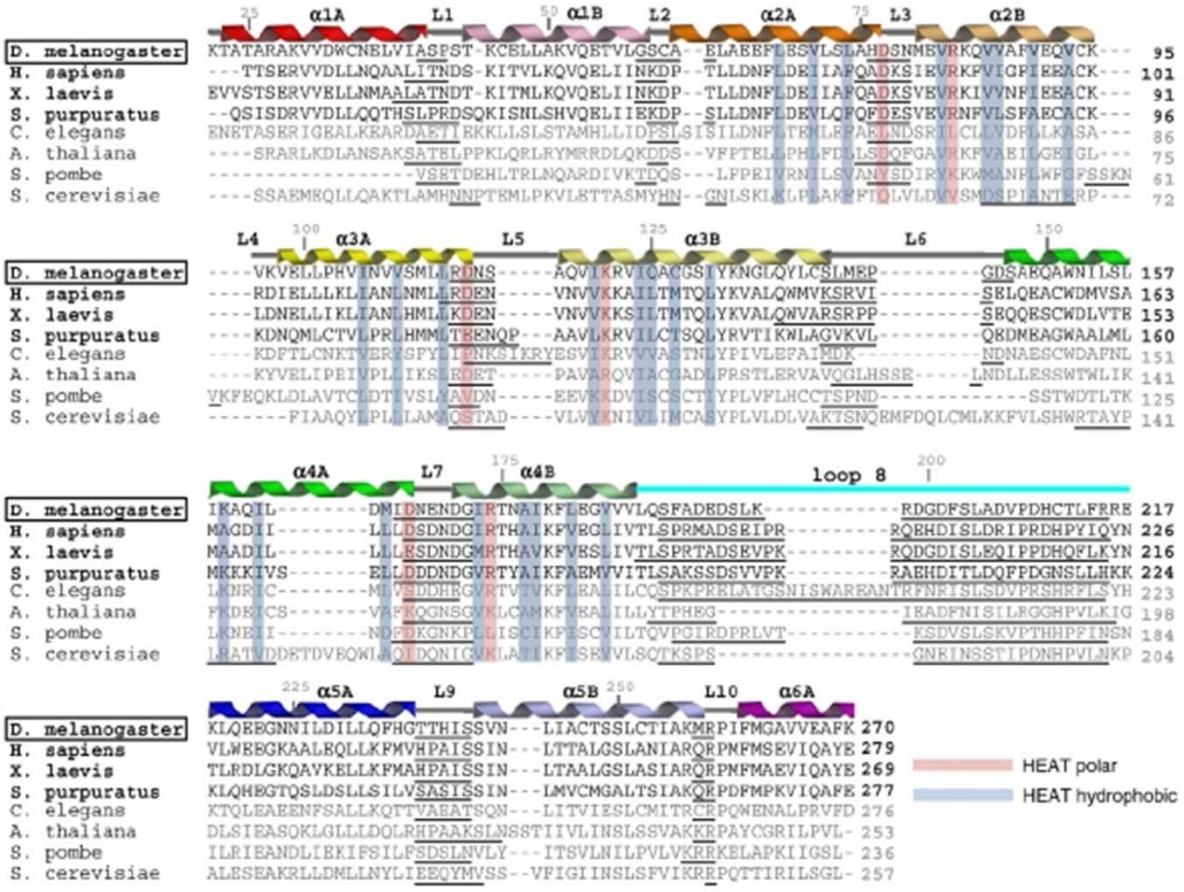

α1A
L1
α1B
L2
α2A
L3
α2B
L4
α3A
L5
α3B
L6
α4A
L7
α4B
loop 8
α5A
L9
α5B
L10
α6A
D. melanogaster
H. sapiens
X. laevis
S. purpuratus
C. elegans
A. thaliana
S. pombe
S. cerevisiae
HEAT polar
HEAT hydrophobic

Figure 3.4

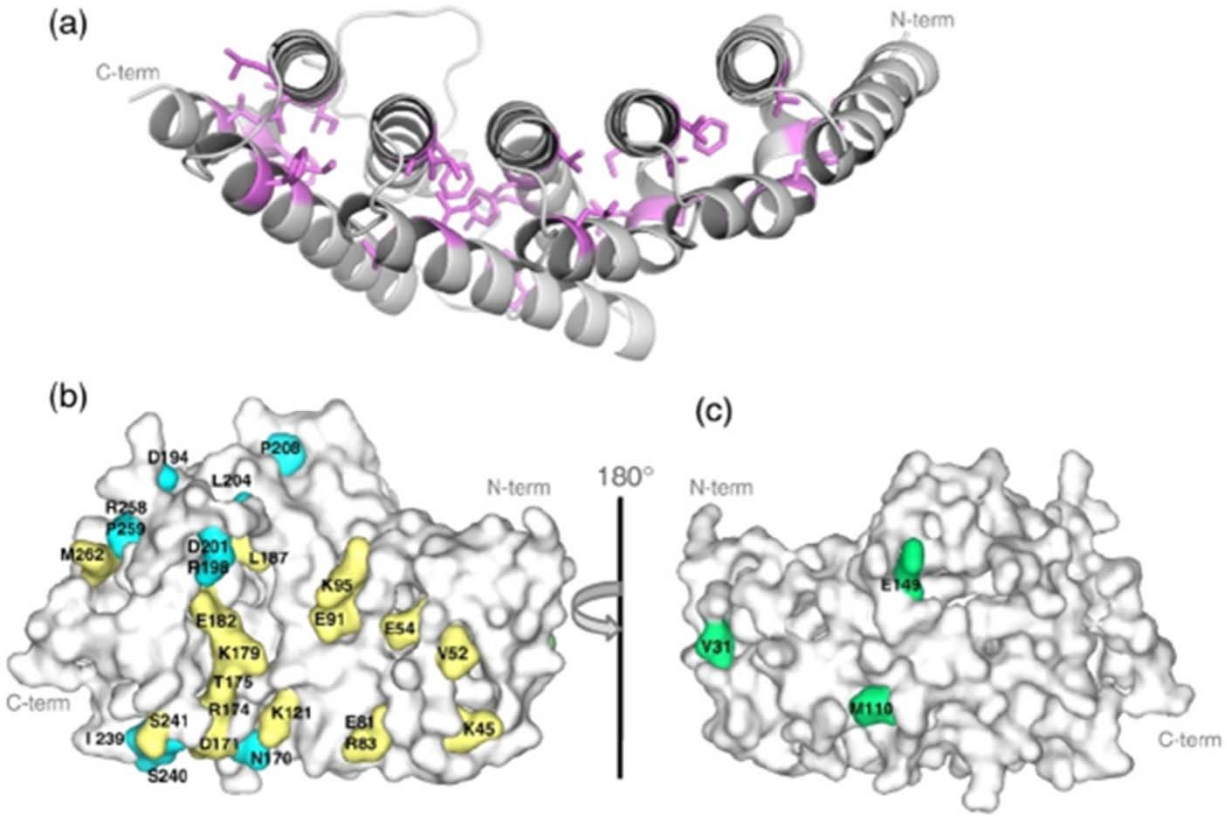

(a)
C-term
N-term
(b)
D194
P208
L204
N-term
180°
R258
P259
M262
D201
R198
L187
K95
E91
E54
V52
E182
K179
T175
R174
K121
E81
R83
K45
C-term
S241
I 239
D171
N170
S240
(c)
N-term
E149
V31
M110
C-term

Figure 3.5

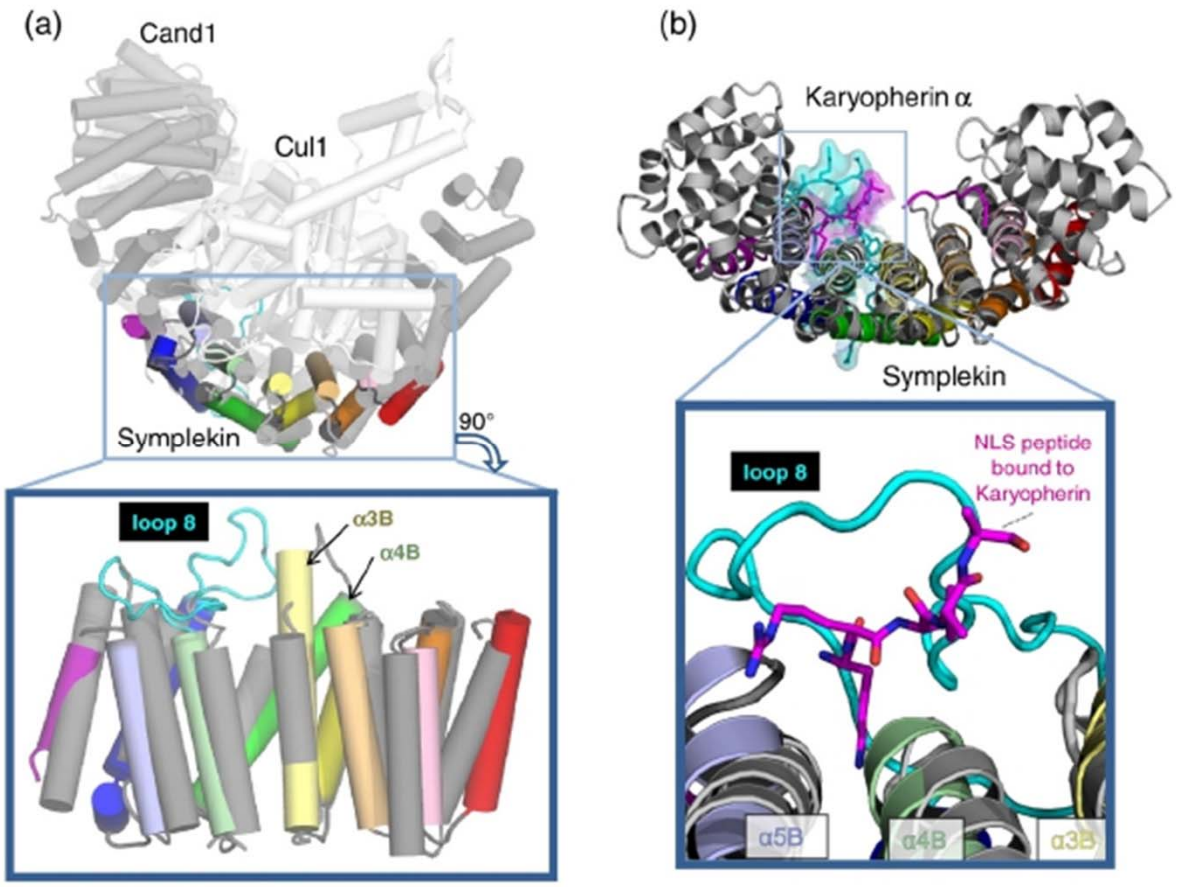

(a)
Cand1
Cul1
Symplekin
90°
loop 8
α3B
α4B
(b)
Karyopherin α
Symplekin
loop 8
NLS peptide
bound to
Karyopherin
α5B
α4B
α3B

Figure 3.6

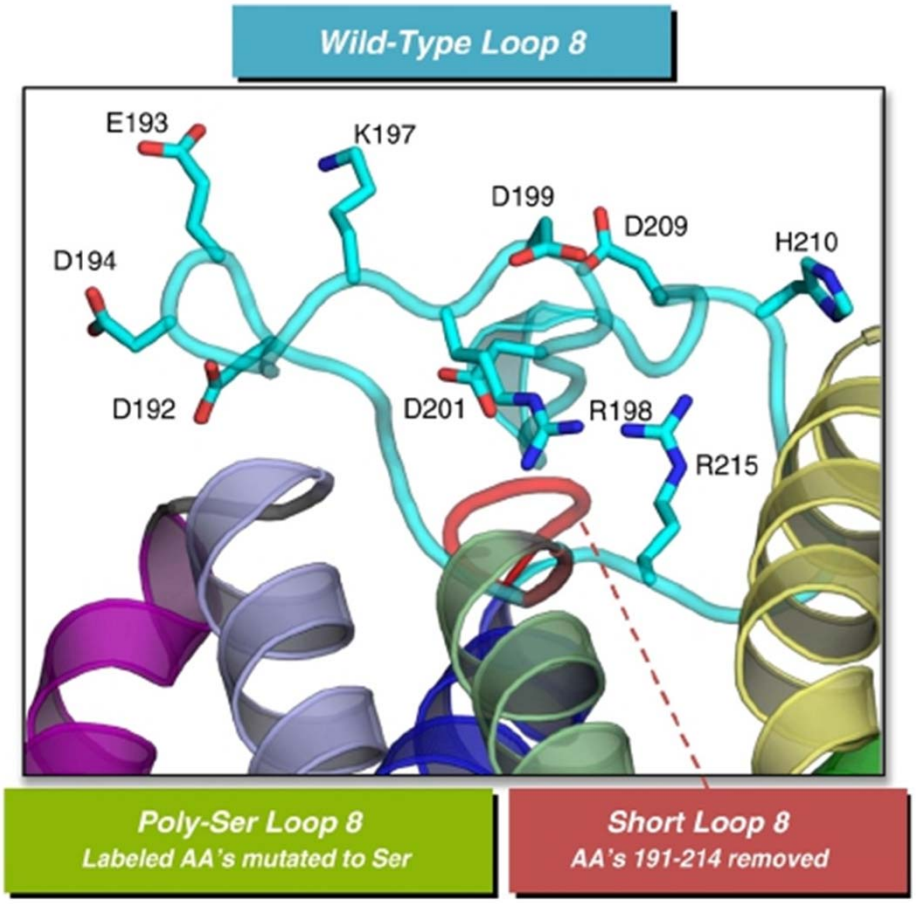

Wild-Type Loop 8
E193
K197
D199
D209
H210
D194
D192
D201
R198
R215
Poly-Ser Loop 8
Labeled AA's mutated to Ser
Short Loop 8
AA's 191-214 removed

Figure 3.7

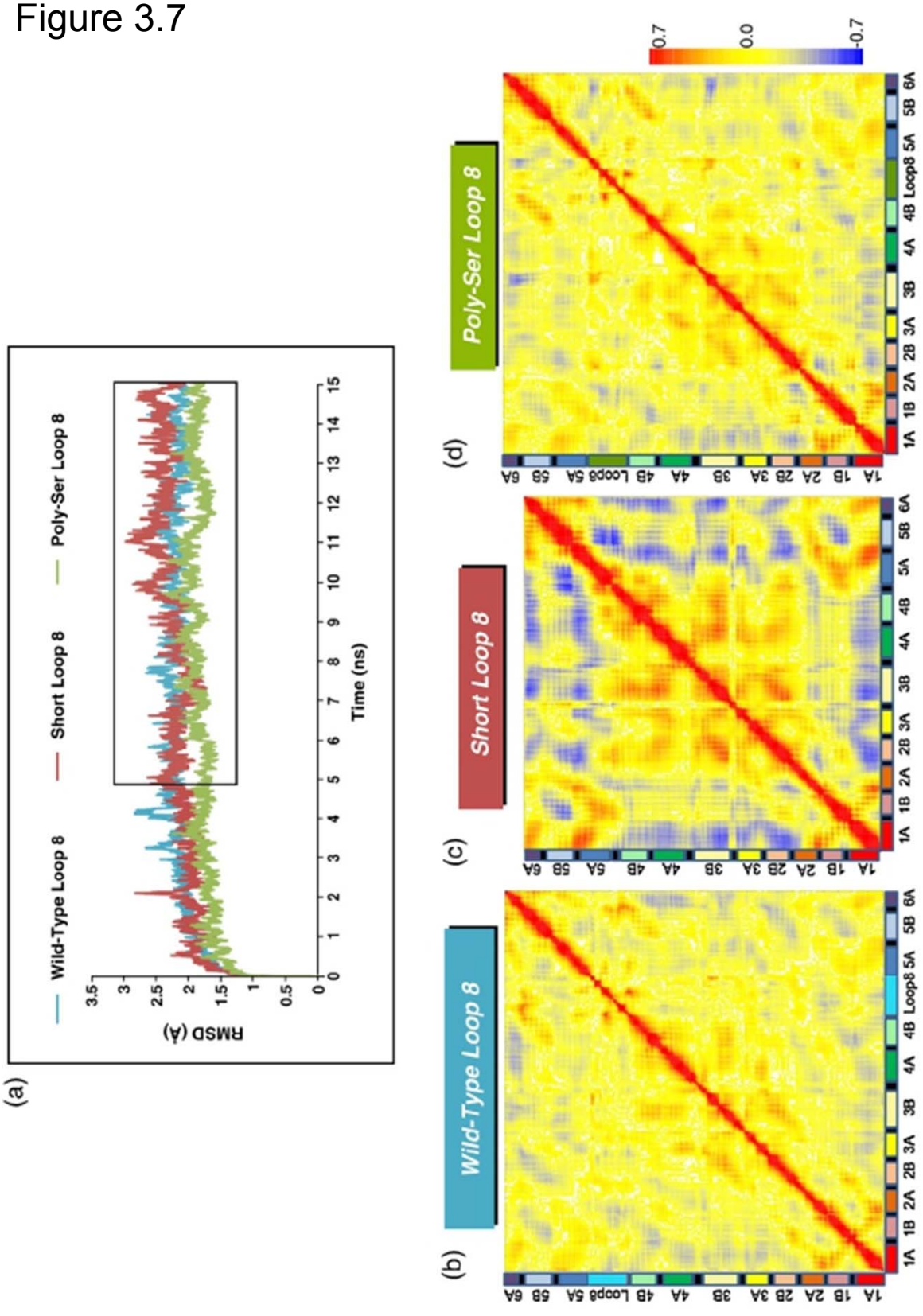
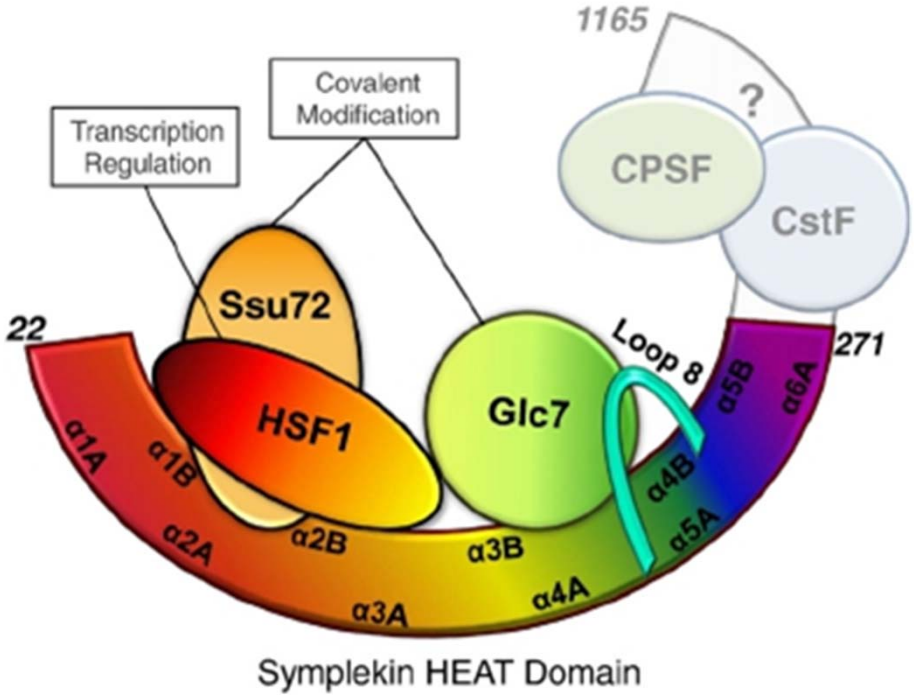

(a)
Wild-Type Loop 8
Short Loop 8
Poly-Ser Loop 8
RMSD (Å)
Time (ns)
(b)
Wild-Type Loop 8
(c)
Short Loop 8
(d)
Poly-Ser Loop 8
0.7
0.0
-0.7

Figure 3.8


1165
?
Covalent Modification
Transcription Regulation
CPSF
CstF
Ssu72
22
Loop 8
271
HSF1
Glc7
α1A
α1B
α2A
α2B
α3A
α3B
α4A
α4B
α5A
α5B
α6A
Symplekin HEAT Domain

3.8 REFERENCES

1. Preiss, T., and Hentze, M. W. (1998) Dual function of the messenger RNA cap structure in poly(A)-tail-promoted translation in yeast, *Nature 392*, 516-520.

2. Sachs, A. B., Sarnow, P., and Hentze, M. W. (1997) Starting at the beginning, middle, and end: translation initiation in eukaryotes, *Cell 89*, 831-838.

3. Mandel, C. R., Bai, Y., and Tong, L. (2008) Protein factors in pre-mRNA 3'-end processing, *Cell Mol Life Sci 65*, 1099-1122.

4. Wilusz, C. J., Wormington, M., and Peltz, S. W. (2001) The cap-to-tail guide to mRNA turnover, *Nat Rev Mol Cell Biol 2*, 237-246.

5. Dominski, Z., and Marzluff, W. F. (2007) Formation of the 3' end of histone mRNA: getting closer to the end, *Gene 396*, 373-390.

6. Takagaki, Y., and Manley, J. L. (2000) Complex protein interactions within the human polyadenylation machinery identify a novel component, *Mol Cell Biol 20*, 1515-1525.

7. Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates III, J. R., Frank, J., and Manley, J. L. (2009) Molecular Architecture of the Human Pre-mRNA 3' Processing Complex, *Mol Cell 33*, 365-376.

8. Sullivan, K. D., Steiniger, M., and Marzluff, W. F. (2009) A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs, *Mol Cell 34*, 322-332.

9. Kolev, N. G., and Steitz, J. A. (2005) Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs, *Genes Dev 19*, 2583-2592.

10. Wagner, E. J., Burch, B. D., Godfrey, A. C., Salzler, H. R., Duronio, R. J., and Marzluff, W. F. (2007) A genome-wide RNA interference screen reveals that variant histones are necessary for replication-dependent histone pre-mRNA processing, *Mol Cell 28*, 692-699.

11. Ghazy, M., He, X., Singh, B. N., Hampsey, M., and Moore, C. (2009) The essential N-terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3'-end processing, *Mol Cell Biol*.

12. Xing, H., Mayhew, C. N., Cullen, K. E., Park-Sarge, O. K., and Sarge, K. D. (2004) HSF1 modulation of Hsp70 mRNA polyadenylation via interaction with symplekin, *J Biol Chem 279*, 10551-10555.

13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res 25*, 3389-3402.

14. Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J., and Kelley, L. A. (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre, *Proteins 70*, 611-625.

15. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. (1998) JPred: a consensus secondary structure prediction server, *Bioinformatics 14*, 892-893.

16. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2006) Pfam: clans, web tools and services, *Nucleic Acids Res 34*, D247-251.

17. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J Mol Biol 299*, 499-520.

18. Rost, B., Yachdav, G., and Liu, J. (2004) The PredictProtein server, *Nucleic Acids Res 32*, W321-326.

19. Zdobnov, E. M., and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro, *Bioinformatics 17*, 847-848.

20. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W., and Bork, P. (2001) Comparison of ARM and HEAT protein repeats, *J Mol Biol 309*, 1-18.

21. Cho, U. S., and Xu, W. (2007) Crystal structure of a protein phosphatase 2A heterotrimeric holoenzyme, *Nature 445*, 53-57.

22. Goldenberg, S. J., Cascio, T. C., Shumway, S. D., Garbutt, K. C., Liu, J., Xiong, Y., and Zheng, N. (2004) Structure of the Cand1-Cul1-Roc1 complex reveals regulatory mechanisms for the assembly of the multisubunit cullin-dependent ubiquitin ligases, *Cell 119*, 517-528.

23. Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A., and Barford, D. (1999) The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs, *Cell 96*, 99-110.

24. Lee, S. J., Matsuura, Y., Liu, S. M., and Stewart, M. (2005) Structural basis for nuclear import complex dissociation by RanGTP, *Nature 435*, 693-696.

25. Matsuura, Y., and Stewart, M. (2005) Nup50/Npap60 function in nuclear protein import complex disassembly and importin recycling, *Embo J 24*, 3681-3689.

26. Paffenholz, R., Kuhn, C., Grund, C., Stehr, S., and Franke, W. W. (1999) The arm-repeat protein NPRAP (neurojungin) is a constituent of the plaques of the outer limiting zone in the retina, defining a novel type of adhering junction, *Exp Cell Res 250*, 452-464.

27. Sampietro, J., Dahlberg, C. L., Cho, U. S., Hinds, T. R., Kimelman, D., and Xu, W. (2006) Crystal structure of a beta-catenin/BCL9/Tcf4 complex, *Mol Cell 24*, 293-300.

28. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn,

D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2008) InterPro: the integrative protein signature database, *Nucleic Acids Res*.

29. Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates, *J Mol Biol 298*, 521-537.

30. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res 25*, 4876-4882.

31. Kelley, L. A., and Sternberg, M. J. (2009) Protein structure prediction on the Web: a case study using the Phyre server, *Nat Protoc 4*, 363-371.

32. Holm, L., and Sander, C. (1996) Mapping the protein universe, *Science 273*, 595-603.

33. Xu, Y., Xing, Y., Chen, Y., Chao, Y., Lin, Z., Fan, E., Yu, J. W., Strack, S., Jeffrey, P. D., and Shi, Y. (2006) Structure of the protein phosphatase 2A holoenzyme, *Cell 127*, 1239-1251.

34. Wang, X., McLachlan, J., Zamore, P. D., and Hall, T. M. (2002) Modular recognition of RNA by a human pumilio-homology domain, *Cell 110*, 501-512.

35. Neuwald, A. F., and Hirano, T. (2000) HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions, *Genome Res 10*, 1445-1452.

36. Conti, E., and Kuriyan, J. (2000) Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha, *Structure 8*, 329-338.

37. Fang, X., Chen, T., Tran, K., and Parker, C. S. (2001) Developmental regulation of the heat shock response by nuclear transport factor karyopherin-alpha3, *Development 128*, 3349-3358.

38. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A., and Pluckthun, A. (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core, *J Mol Biol 376*, 1282-1304.

39. Zachariae, U., and Grubmuller, H. (2006) A highly strained nuclear conformation of the exportin Cse1p revealed by molecular dynamics simulations, *Structure 14*, 1469-1478.

40. Zachariae, U., and Grubmuller, H. (2008) Importin-beta: structural and dynamic determinants of a molecular spring, *Structure 16*, 906-915.

41. Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublie, S. (2008) Crystal structure of the 25 kDa subunit of human cleavage factor Im, *Nucleic Acids Res 36*, 3474-3483.

42. Qu, X., Perez-Canadillas, J. M., Agrawal, S., De Baecke, J., Cheng, H., Varani, G., and Moore, C. (2007) The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA 3'-end processing, *J Biol Chem 282*, 2101-2115.

43. Balbo, P. B., Meinke, G., and Bohm, A. (2005) Kinetic studies of yeast polyA polymerase indicate an induced fit mechanism for nucleotide specificity, *Biochemistry 44*, 7777-7786.

44. Deo, R. C., Bonanno, J. B., Sonenberg, N., and Burley, S. K. (1999) Recognition of polyadenylate RNA by the poly(A)-binding protein, *Cell 98*, 835-845.

45. Perez-Canadillas, J. M. (2006) Grabbing the message: structural basis of mRNA 3'UTR recognition by Hrp1, *Embo J 25*, 3167-3178.

46. Mandel, C. R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L., and Tong, L. (2006) Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease, *Nature 444*, 953-956.

47. Bai, Y., Auperin, T. C., Chou, C. Y., Chang, G. G., Manley, J. L., and Tong, L. (2007) Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors, *Mol Cell 25*, 863-875.

48. Meinhart, A., and Cramer, P. (2004) Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors, *Nature 430*, 223-226.

49. Noble, C. G., Beuth, B., and Taylor, I. A. (2007) Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor, *Nucleic Acids Res 35*, 87-99.

50. Isgro, T. A., and Schulten, K. (2005) Binding dynamics of isolated nucleoporin repeat regions to importin-beta, *Structure 13*, 1869-1879.

51. He, X., and Moore, C. (2005) Regulation of yeast mRNA 3' end processing by phosphorylation, *Mol Cell 19*, 619-629.

52. Rizzuti, B., Sportelli, L., and Guzzi, R. (2001) Evidence of reduced flexibility in disulfide bridge-depleted azurin: a molecular dynamics simulation study, *Biophys Chem 94*, 107-120.

53. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs, *Nucleic Acids Res 31*, 3635-3641.

54. Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, *Bioinformatics 21*, 3369-3376.

55. Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences, *Science 252*, 1162-1164.

56. Donnelly, M. I., Zhou, M., Millard, C. S., Clancy, S., Stols, L., Eschenfeldt, W. H., Collart, F. R., and Joachimiak, A. (2006) An expression vector tailored for large-scale, high-throughput purification of recombinant proteins, *Protein Expr Purif 47*, 446-454.

57. Otwinowski, Z. a. M., W. (1997) *Processing of X-ray Diffraction Data Collected in Oscillation Mode, in Methods in Enzymology, Macromolecular Crystallography, Part A*, Vol. 276, Academic Press, New York.

58. Fu, Z. Q., Rose, J., and Wang, B. C. (2005) SGXPro: a parallel workflow engine enabling optimization of program performance and automation of structure determination, *Acta Crystallogr D Biol Crystallogr 61*, 951-959.

59. Emsley, P. a. C., Kevin. (2004) Coot: Model-Building Tools for Molecular Graphics, *Acta Crystallographica Section D - Biological Crystallography 60*, 2126-2132.

60. Collaborative computational project. (1994) The CCP4 Suite: Programs for Protein Crystallography, *Acta Cryst. D50*, 760-763.

61. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation, *Proteins 50*, 437-450.

62. Delano, W. L. (2002) The PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA, USA.

63. Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs, *J Comput Chem 26*, 1668-1688.

64. Jorgensen, W., Chandrasekhar J, Madura JD, Impey RW, Klein ML. (1983) Comparison of simple potential functions for simulating liquid water., *J Chem Phys 79*, 926-935.

65. Essman U, P. L., Berkowitz ML, Darden T, Lee H, Pedersen L. (1995) A smooth particl mesh Ewald method, *J Chem Phys 103*, 8577-8593.

66. Teotico, D. G., Frazier, M. L., Ding, F., Dokholyan, N. V., Temple, B. R., and Redinbo, M. R. (2008) Active nuclear receptors exhibit highly correlated AF-2 domain motions, *PLoS Comput Biol 4*, e1000111.

CHAPTER 4

Active Nuclear Receptors Exhibit Highly Correlated AF-2 Domain Motions

4.1 INTRODUCTION

The nuclear receptor (NR) superfamily of ligand-regulated transcription factors controls the expression of genes essential to metabolism, development and systemic homeostasis [1,2,3]. NRs are modular proteins typically composed of a conserved N-terminal Zn-module DNA binding domain (DBD) that targets specific response elements, a variable hinge region, and a C-terminal ligand binding domain (LBD) capable in most cases of responding to specific small molecule ligands [4]. NR LBDs contain a shallow activation function 2 (AF-2) surface formed by helices α3, α3', α4 and αAF that is essential for ligand-dependent interactions with transcriptional coregulators. The AF-2 surface complexes with LxxLL-containing transcriptional coactivators in the presence of agonist ligands, and with distinct leucine-rich corepressor motifs in the presence of antagonists or in the absence of ligand [4,5].

The pregnane X receptor (PXR) controls the expression of a wide range of gene products involved in xenobiotic metabolism and endobiotic homeostasis [6,7,8], and is

Teotico DG, Frazier ML, Ding F, Dohkolyan NV, Temple BR, Redinbo MR. Active nuclear receptors exhibit highly correlated AF-2 domain motions. *PLoS Comput Biol.* 2008 Jul 1;4(7):e1000111.

Monica Frazier contributed Section 4.3.1, Figures 4.2, 4.3A, Supplemental Figures 4.1, 4.2, 4.3, and 4.4, and oversaw the revision process, including re-running and/or extending the length of all MD simulations mentioned in the text.

unusual in the NR superfamily in several respects. First, PXR responds promiscuously to a wide range of chemically-distinct ligands from small lipophilic phenobarbital (232 Da) to the large macrolide antibiotic rifampicin (823 Da); in contrast, most NRs are highly specific for their cognate ligands [9,10,11]. Second, the PXRs of known sequence contain a 50-60 residue insert that, as observed in human [12,13,14], creates a unique β-turn-β motif and novel PXR homodimer interface. All NR LBDs fold into a three-layer α-helical sandwich in which α10 forms standard homodimerization interactions (*e.g.*, for steroid receptors like the estrogen receptor-α, ERα) or heterodimerization interactions (*e.g.*, with RXR for orphan receptors like PXR) [2,15,16]. The PXR LBD, in contrast, contains a second oligomerization interface at the novel β-turn-β motif in which intercalating tryptophan and tyrosine residues (Trp-223/Tyr-225) lock across the dimer to form an aromatic zipper [4,5,12] (Figure 4.1A). It has been shown that this dimer interface is essential to PXR function, and that the specific disruption of homodimerization eliminates the ability of the receptor to interact with transcriptional coactivators like steroid receptor coactivator 1 (SRC-1), but does not impact PXR's subcellular localization or its association with DNA, RXR, or activating ligands [12]. This work led to the proposal of a PXR-RXR heterotetramer as the functional unit [12] (Figures 4.1A, 4.1B).

The unique PXR homodimer interface, however, is located more than 30 Å from the coactivator binding site at the receptor's AF-2 surface (Figure 4.1A). Thus, we hypothesize that long-range motions within the PXR LBD are essential for communicating the stabilizing effect of PXR homodimerization to the AF-2 domain. To test this hypothesis, we performed all-atom molecular dynamics (MD) simulations on both the PXR LBD, as well as two other nuclear receptor LBDs, in various states (Table 4.1). The former orphan peroxisome proliferator-activated receptor-γ (PPARγ) is functional as a heterodimer with RXR, while the steroid estrogen receptor-α (ERα) is active as an analogous homodimer (Figure 4.1B). We examined LBDs in inactive states (*e.g.,* monomers or mutants), as well as those in the

proper functional states (*e.g.,* homo- or heterodimers, or as a heterotetramer for RXR) we have termed “active-capable.” Our results support the conclusion that the NR LBD provides a scaffold for long-range motions that prepare the AF-2 surface for binding to transcriptional coactivators.

4.2 RESULTS

4.2.1 Stable Dynamic Trajectories

Six all-atom molecular dynamics (MD) runs were performed for 20-30 ns on three nuclear receptor LBDs (Table 4.1). PXR was examined both as a heterodimer with RXR and as a heterotetramer with RXR (30 ns simulations). Wild-type PPARγ was examined as a heterodimer with RXR, and the inactive PPARγP467L mutant was also examined as a heterodimer with RXR (20 ns simulations). Finally, ERα was examined both in its inactive monomeric state (20 ns), and as an ERα homodimer (25 ns). All six trajectories were judged as stable by two criteria. First, the total energy of each system, calculated as the sum of kinetic and potential energy at each time point, was found to be essentially constant after the first 2-3 ns (Figure 4.2, Supplemental Figure 4.1). These results indicate that after a short period of equilibration, each simulation was sampling an energetically stable conformational ensemble. Second, all trajectories were analyzed in terms of moving average all-atom root mean square deviations (RMSDs) from starting crystal structures over the simulation time course (Supplemental Figures 4.2, 4.3). The PXR-RXR trajectories exhibited RMSD values of 0.7-5.0 Å (Supplemental Figure 4.2), while the PPARγ- and ERα-containing trajectories exhibited values of 1.7-3.5 Å (Supplemental Figure 4.3). Such deviations were considered low for systems of this size (*e.g.,* 1044 residues for the PXR-RXR heterotetramer). The RMSD results indicate that all simulations were stable for at least

the last 10 ns of each trajectory (Supplemental Figures 4.2, 4.3). Thus, the final 10 ns section of each simulation was used for subsequent analysis.

4.2.2 Highly Correlated Motion in the PXR-RXR Heterotetramer

The PXR-RXR heterotetramer is expected to have distinct functional dynamics relative to the heterodimer because the heterotetramer contains the unique PXR homodimer interface shown to be essential for receptor activity [12] (Figure 4.1). Thus, we examined the PXR LBDs in both the PXR-RXR heterodimer and PXR-RXR heterotetramer simulations over the last 10 ns of each trajectory using essential dynamics analysis. Essential dynamics discriminates between concerted motions of residue clusters within a protein and uncorrelated residue fluctuations [17]. We computed normalized covariance matrices [18] to classify the relationships between all possible residue pairs in the protein (Figure 4.3A). In this analysis, correlation (two residues moving in the same direction) is indicated by residue-residue correlation coefficients approaching +1, while correlation coefficients approaching -1 indicate anticorrelation (residues moving in opposite directions). Correlation coefficients near zero, in contrast, are associated with residue pairs that lack a dynamic relationship.

The PXR LBDs in the PXR-RXR heterotetramer exhibit significantly more residue-residue correlation relative to the PXR LBD in the PXR-RXR heterodimer (Figure 4.3A). Indeed, the distribution of correlation coefficients for the PXR LBD in the PXR-RXR heterodimer has one peak centered close to zero, indicating the majority of residue-pairs are not correlated (data not shown). In contrast, the correlation coefficient distribution for the PXR LBDs in the PXR-RXR heterotetramer has two distinct peaks, one positive and one negative, indicating both residue-residue correlation and anticorrelation (data not shown).

Clusters of correlated PXR residues from the PXR-RXR heterodimer and heterotetramer that exhibited concerted motion were then examined for the strength of their residue-residue correlation coefficients and the biological significance of those dynamics.

Clustering the PXR LBDs from the PXR-RXR heterotetramer at a correlation coefficient less than 0.6 produced a single cluster containing the complete PXR-RXR homodimer, while clustering at a correlation coefficient above 0.8 resulted in clusters comprised of only 2-5 residues. Neither coefficient cutoff alone could interrogate the biological relevance of the concerted motions; thus, we classified clusters using three correlation coefficient cutoffs (Figure 4.3B, C). Such cutoffs discriminate between weak (0.6 or less), medium (between 0.6 and 0.7), and strong (between 0.7 and 0.8) correlations between PXR residues. The same cutoffs set in the heterotetramer were used, for consistency, to cluster the relatively weak correlated motion observed in the PXR-RXR heterodimer. Indeed, the PXR LBD from the heterodimer exhibited only five small correlated clusters, with smaller regions of these weakly correlated clusters remaining at a medium strength correlation coefficient, and only one group of a few residues identifiable at a strong correlation coefficient (Figure 4.3B).

In distinct contrast, however, the PXR LBDs in the PXR-RXR heterotetramer form a strongly correlated unit (Figure 4.3C). The β-sheet region involved in the PXR-PXR homodimerization interface (β1, β1', β3, β4), together with α-helices 1, 3, 3', 4 and 9, exhibit the strongest degree of correlation; the residues of these β-sheets and α-helices are all clustered together with a correlation coefficient of 0.8. The neighboring helices, including αAF, also exhibit highly correlated motion with correlation coefficients > 0.6 (Figure 4.3C). The strength of residue-residue correlations throughout this region suggest that α3 forms a critical conduit through which the stabilizing effects of the homodimer interface involving β1, β1', β3, and β4 are communicated to helices 3, 3', 4 and AF of the AF-2 surface. In the PXR-RXR heterodimer, however, the same β-sheet region is anticorrelated with the AF-2 domain (Figure 4.3B).

4.2.3 Highly Correlated Motion in the PXR-RXR Heterotetramer AF-2 Surface

We next examined the motions in the four helices that create the AF-2 coactivator binding surface on PXR: α3, α3', α4, and αAF. The concerted motion of this surface was compared between the PXR LBDs in the PXR-RXR heterodimer and heterotetramer trajectories, and was examined using both quasiharmonic analysis (QHA) and normal mode analysis (NMA). Both methods have benefits and limitations. For quasiharmonic analysis, its benefits are all-atom resolution and the use of explicit solvent, but it is limited by the time constraints of all-atom MD. Normal mode analysis has the benefit of observing motions on a longer timescale than available with QHA, but is limited to analyses based upon the coarse grained model solely of the macromolecule. Our results agree with others that it takes more NMA modes than QHA modes to describe the same motions [19]. Thus, we employed the first two modes from QHA and first 14 nontrivial modes from NMA (see Methods). Eigenvectors from these analyses are associated with the magnitude and direction of motion, and these eigenvectors can be used to create visuals of the NR's motion.

After examining the vectors describing the primary modes of motion derived from QHA for each α-carbon position in the PXR LBDs of the PXR-RXR simulations, a single average vector was calculated to describe the motion of seven of the eleven α-helices in the LBD. The remaining four helices, α3, α4, αAF and α10, displayed distinct motions at their termini; thus, for these helices, two average vectors were employed. The results of this analysis show that the PXR LBD helices from the PXR-RXR heterotetramer move as a single unit, and in one direction (Figure 4.4A). This correlation is especially evident in the AF-2 surface, as α3, α3', α4 and αAF all move together in the same direction (Figure 4.4A inset). In contrast, the PXR LBD from the PXR-RXR heterodimer exhibited relatively small, disjointed motions (Figure 4.4B). This lack of helix-helix correlation includes the AF-2 surface helices α3, α3', α4 and αAF (Figure 4.4B inset).

AF-2 mobility identified by QHA was also assessed by examining the angles between the directions of motion as defined by the eigenvectors for α-carbons of residues important to coactivator binding (Table 4.2, Methods). As such, if two residues in the AF-2 surface are moving together, the angle between them is small (see Methods, Equation 1). The average angle from the sum of motion vectors (modes) 1 and 2 between AF-2 domain residues in the PXR-RXR heterodimer simulation was 71.6°. In contrast, the average angle for the same residue pairs in the PXR-RXR heterotetramer simulation was 31.5° (Table 4.2). Taken together, these QHA results support the conclusion that the intramolecular β-sheet formed by the PXR homodimer interface produces highly correlated AF-2 surface motions in the PXR-RXR heterotetramer complex.

In a second analysis, modes of motion of the AF-2 surface of the PXR LBD were examined from both the heterodimer and heterotetramer trajectories using NMA. Similar to the QHA study above, angles between the directions of motion as defined by the eigenvectors for α-carbons of residues important to coactivator binding were calculated (Table 4.2, Methods). The average angle observed in the AF-2 surface in the PXR-RXR heterotetramer was 13.5° using NMA, even smaller than the average angle found using QHA (Table 4.2). In contrast, the average angle for the same PXR AF-2 residues in the PXR-RXR heterodimer was 70.1°, nearly identical to the value found using QHA (Table 4.2). Thus, these data support the conclusions of the QHA study, and indicate that a high degree of helix-helix correlation is present in the AF-2 surface of the PXR-RXR heterotetramer relative to the heterodimer. Similarities between the QHA and NMA results strengthen this collective conclusion, particularly because QHA is based on shorter dynamic movements of all atoms, while NMA examines harmonic oscillations that occur on longer time scales.

Plots of the angles between the vectors of motion of all possible PXR LBD residue pairs from both the heterodimer and heterotetramer simulations for the QHA and NMA studies are shown in Figures 4.5A and 4.5B, respectively. Areas in green represent angle

values close to zero (vectors moving in the same direction, or correlated), while areas in yellow indicate vectors with angles close to 180º (vectors moving in the opposite direction, or anticorrelated). In both plots, a high degree of correlated motion is observed for the PXR LBD in the PXR-RXR heterotetramer, while significantly less correlation is observed for the LBD in the heterodimer (Figures 4.5A, B). The similarity between Figures 5A and B, from selected modes of QHA- and NMA-identified motion, and Figure 4.3A, from all modes of motion, indicates that enough modes were chosen in both QHA and NMA to represent the motion of each LBD (Methods). In addition, both the QHA and NMA plots for the heterotetramer indicate similar correlated structural elements. For example, the PXR β-sheet moves in a more correlated manner with respect to αAF in the heterotetramer relative to the heterodimer (Figures 4.5A, B). In summary, long-range motions impacted by the oligomeric state of PXR play a central role in the function of this nuclear xenobiotic receptor.

4.2.4 Correlated AF-2 Motions in Other Nuclear Receptors

We next examined whether the unliganded LBDs of other members of the NR superfamily would also exhibit correlated AF-2 surface motions. As stated above, 20-25 ns MD simulations were performed on two inactive NR states, the ERα monomer and the PPARγ P467L-RXR heterodimer complex, and on two "active-capable" states, the ERα homodimer and the wild-type PPARγ-RXR heterodimer. A P467L mutation has been shown to inactivate PPARγ [20]. Only moderate levels of residue-residue correlation and anticorrelation were observed for both states of ERα and PPARγ (Supplemental Figures 4.4A, B). Examination of correlation coefficient distributions in these simulations reveals that all remain close to zero, indicating relatively non-correlated motion (data not shown).

In spite of their relatively limited overall correlation, however, the active-capable forms of ERα and PPARγ-RXR exhibited correlated AF-2 domain motions. Similar to the analysis of the PXR trajectories, both QHA and NMA were employed to examine these ERα

and PPARγ simulations. Results from QHA studies reveal that the active-capable forms of ERα, and PPARγ exhibit more correlated AF-2 motions than their inactive counterparts (Figure 4.6). Angles between the vectors describing AF-2 surface helix motions in PPARγ and ERα states using both QHA and NMA further support the overall conclusion that active-capable states exhibit correlated AF-2 surfaces (Table 4.3, 4.4). For example, the average angles for ERα homodimer and wild type PPARγ-RXR determined using NMA are 41.0° and 48.8°, respectively, while those for the inactive ERα monomer and the PPARγ P467L mutant are 63.1° and 58.3°. Again, the AF-2 correlation in motion observed using the shorter time scales of all-atom molecular dynamics simulations and QHA are also seen in the longer harmonic oscillations of NMA. In summary, correlated motion appears to be a consistent feature in the AF-2 domains of active-capable nuclear receptor LBDs.

4.3 DISCUSSION

The differences in human PXR LBD motion between two oligomeric states of the receptor (as a heterodimer and a heterotetramer with RXR) were examined using molecular dynamics trajectories, essential dynamics, quasiharmonic, and normal mode analyses. It was hypothesized that the PXR heterotetramer, in which PXR LBD monomers form a unique homodimer shown to be critical for transcriptional regulation [12], would exhibit functionally-relevant motion. Indeed, we find that this “active-capable” form of PXR exhibits not only significantly more overall motion and more correlated motion relative to the heterodimer, but also highly correlated motion in the AF-2 surface responsible for functionally-essential contacts with transcriptional coactivators (Figures 4.4,4.5). These data suggest that a high degree of motion promotes the proper function of this nuclear receptor, provided that the motion is correlated to preserve the state of the receptor ready to bind to leucine-rich coactivator motifs.

In addition, these results indicate that long-range motions are critical to the function of the xenobiotic receptor PXR. The homodimer interface unique to the PXR LBD is located approximately 30-35 Å from the AF-2 surface (Figure 4.1). Essential dynamics have revealed that the β-sheet and six α-helices in PXR (1, 3, 3', 4, 9, AF), including those that comprise the AF-2 surface, move as a single unit in the heterotetramer trajectory (Figure 4.3). This suggests a structural mechanism by which PXR homodimerization creates a ten-stranded intermolecular β-sheet (Figure 4.1) that positively impacts AF-2 domain motion. The N-terminal portion of α3 appears to serve as a critical bridge between the PXR β-sheet and the AF-2 helices, such that correlated α1-α4 motion is "communicated" to α3-α4 and αAF (Figure 4.4). This relationship explains how the obligate PXR monomer mutant Trp-223-Ala/Tyr-225-Ala, in which the interlocking aromatic residues at the homodimer interface are eliminated, is still able to bind to ligand, DNA and RXR, but not to transcriptional coactivators at the AF-2 surface [12].

This hypothesized path of "communication through motion" mediated by α3 and involving several β-strands, as well as α1 and α 9, correlates well with existing PXR structure-function data. First, Met-243, located in the N-terminal portion of α 3, is contacted by ligands in all reported PXR LBD crystal structures [4,21,22]. Thus, they appear critical for the ligand-enhanced transcriptional activity exhibited by PXR. Second, single mutations in either α3 or α3', such as Thr-248-Glu, Lys-277-Gln and Pro-268-His, result in a loss of PXR activity [23,24]. In addition, although the α3 double-mutant Lys-277-Gln/Thr-248-Glu restores transcriptional activation, it abolishes the antagonism of ketoconazole, hypothesized to function by binding the AF-2 surface [24,25]. Third, the α1 and α9 mutants Asp-163-Gly and Ala-370-Thr, respectively, represent a class of PXR variants that are distantly located from the AF-2 domain but result in reduced transcriptional activity [26]. Taken together, these data support the conclusion that the wild-type PXR LBD is "tuned" in

its heterotetrameric complex with the RXR LBD to produce correlated motions that promote the binding of transcriptional coactivators.

Extension of this analysis into other nuclear receptors reveals correlated AF-2 surface motions in “active-capable” forms of ERα and PPARγ (Figure 4.6, Supplemental Figure 4.4). Thus, long-range motions may play critical roles in the LBD activation potential of several members of the nuclear receptor superfamily. Our results expand on previous MD investigations of NR LBDs. For example, dynamics studies on ERα [27] showed that the addition of coactivator peptide and ligand to apo ERα lead to increased αAF helix motion in unspecified directions. Similarly, studies on androgen insensitivity syndrome associated androgen receptor Pro-892-Ala and Pro-892-Leu mutations revealed via biochemical assays and MD simulations an increased flexibility and distortion of the αAF helix [28]. We present evidence that the AF-2 domain helices of the Erα, PPARγ, and PXR LBDs move together and in the same direction in each receptor. One may postulate that the uncorrelated motion between the helices in the AF-2 domain observed for inactive receptors (e.g., apo PXR-RXR heterodimer, ERα monomer and the PPARγ P467L-RXR mutant) may represent the initial transition towards an αAF position required for corepressor binding [29]. Alternatively, these anticorrelated motions may simply prevent coactivator binding to LBDs that are not in active-capable oligomeric states.

The results presented here are also in agreement with limited proteolysis [15], fluorescence polarization [20], and NMR [30,31] studies that examined the stabilization of global and local motions of ERα [15] and PPARγ [20] upon ligand binding. Of particular note are time-resolved fluorescence polarization studies by Kallenberger and Schwabe [20] on the human P467L PPARγ mutant that causes insulin resistance and early onset hypertension. This mutation was found to weaken immobilization of αAF against the main body of the receptor. In our molecular dynamics simulations, wild type PPARγ-RXR exhibited a strong degree of correlated AF-2 motion while the PPARγ P467L-RXR mutant

showed uncorrelated motion in its AF-2 domain (Figure 4.6). This is the first model of nuclear receptor dynamics that relates changes in motion to a mutation causing a disease state.

While nuclear receptors are well-established targets for small molecule modulators that treat a wide range of conditions, current drugs function as agonists and antagonists via the ligand binding pocket. However, recent data have indicated that nuclear receptor LBDs can be antagonized using small molecules that block coregulator binding to the AF-2 surface. For example, thyroid receptor antagonists discovered by high-throughput screening were found to act at the AF-2 site of that receptor [32,33]. In addition, the azole family of antifungal compounds has recently been shown to antagonize the action of human PXR via the AF-2 domain [24,25]. The dynamics data presented here further elucidate the nature of motions essential for AF-2 active-capable function, and may facilitate the improved design or development of therapeutics targeted to specific NR AF-2 surfaces.

4.4 METHODS

Molecular Dynamics Simulations

Molecular dynamics simulations were run on the apo PXR-RXR LBD heterodimer and heterotetramer. MD simulations were also performed for the nuclear receptors ERα (monomer and homodimer) and PPARγ (wild-type heterodimer with RXR and mutant P467L heterodimer with RXR). A summary of these simulations containing their oligomeric states, starting structure PDB IDs, and activity is provided in Table 4.1. All starting structures were obtained from the protein databank (www.rcsb.org). The PXR-RXR heterodimer and heterotetramer models as proposed in Noble et. al. [12] were generated by first generating a PXR-RXR heterodimer model, followed by overlaying two copies of the heterodimer onto each protomer of the PXR homodimer structure. The PXR-RXR heterodimer model was created by superimposing the PXR LBD onto the LBD of PPARγ in the PPARγ-RXRα

heterodimer crystal structure (PDBID: 1FM6). Upon creating this model, the PXR LBD was found to make nearly identical salt bridges, hydrogen bonds and hydrophobic interactions with the RXRα LBD as seen in the PPARγ-RXRα heterodimer crystal structure.

All MD simulations were carried out with a 2 fs time step using the AMBER 2003 force field [34]. Molecular graphics figures were generated in Pymol (http://pymol.sourceforge.net). All production runs employed the PMEMD module from Amber 9.0 [35]. Frames were recorded every 0.4 ps. Topology and parameter files were created using the LEaP program within AMBER [35]. The simulation system consisted of the protein surrounded by a truncated octahedron of water and sodium ions to maintain charge neutrality. An explicit solvent model was used with TIP3P water molecules filling 12.5 Å between the surface of each protein and the edge of the box [36]. Electrostatic interactions were calculated using the particle-mesh Ewald algorithm [37] with a cutoff of 10 Å applied to Lennard-Jones interactions.

The SANDER package within AMBER was used for 5000 steps of energy minimization. Equilibration included 20 ps of constant volume conditions with heating from 100 to 300 K followed by 100 ps constant temperature conditions. Constant volume heating from 200 to 300 K was applied to the system for 20 ps before beginning the production run with the NPT ensemble.

Simulations were analyzed using the PTRAJ package in Amber [35]. All-atom moving average root-mean-square deviations (RMSD) were calculated for each trajectory using the initial crystal structure as reference with an interval of 100 data points. Quasiharmonic analysis was employed for each trajectory using PTRAJ [35].

Loop Modeling

In all PXR simulations, a disordered loop (PDB ID 1ILG, residues 178-197) missing from the apo PXR LBD crystal structure was modeled using the MODELLER module of InsightII with database searching (www.accelrys.com) [38]. The N and C termini of the

modeled loop segment were reconnected to the missing sections of the crystal structure to avoid the termini from unrealistic flopping during simulations. The loop was examined for its potential impact on the RMSD from starting crystal structure by analyzing the simulations of the PXR LBD with and without the loop. The loop was found to impact the overall magnitude, but not the variability of the RMSD, suggesting that these regions move more than others, but do not effect stable conformations sampled during the simulation. Therefore, we have omitted the loop from subsequent analyses. However, we chose to include this loop in our simulations because it is a more realistic biological representation of the receptor.

Correlation Analysis

The pair-wise correlation coefficient as described in Sharma et. al. [18], C_{ij}, was computed between α-carbons of two residues, i and j, with values ranging from -1 to +1. The more positive the value of C_{ij}, the more correlated (moving in the same direction with one another) the two residues, i and j, move. Likewise, the more negative the value of C_{ij}, the more anticorrelated (moving in the opposite direction to one another) the two residues, i and j, move. The single-linkage clustering method [39] was applied to identify distinct sets of residues that move correlated with each other or anticorrelated to each other. In this method, a graph is initially built where each entity corresponds to individual residues. The clustering method proceeds by first finding two entities that have the highest similarity (i.e., the correlation coefficient) between them. After clustering those two entities into one, the similarities between this new entity and the rest are updated. This process is repeated until there are no more entities to cluster or the correlation coefficient cutoff is satisfied. In a single-linkage clustering method, the similarity between two clusters is defined as the largest similarity or the highest correlation coefficient between any two members from the two clusters.

Angle Analysis

Residues chosen to describe motion in Tables 4.2-4 were not chosen at random in the AF-2 domain of PXR LBD. Glu-427 of αAF and Lys-259 of α3 are the "charge clamp" residues of PXR; the charge clamp is a common structural motif in nuclear receptor-coactivator interactions and involves contacts between the LBD and the termini of the coactivator LxxLL helix. Lys-277 of α4 was chosen because it is conserved in many receptors and Leu-424 of αAF directly contacts the coactivator SRC-1 [14]. Angle analysis was performed using Equation 1 to find the average angle between vectors for α-carbon a and b.

$$\theta = \cos^{-1}\left[\frac{\left(\vec{a}\cdot\vec{b}\right)}{\left(\left|\vec{a}\right|\left|\vec{b}\right|\right)}\right] \quad (1)$$

Quasiharmonic Analysis

The effective modes of vibrational motion can be obtained using quasiharmonic analysis by calculating a force field relative to the average structure based on the fluctuations generated from an MD simulation. Quasiharmonic modes, unlike standard principal component methods, are mass weighted just as normal modes and thus may be compared directly with normal mode analysis. However in quasiharmonic modes, anharmonic effects are implicitly included and thus may be different from normal modes [40]. The percent contribution of each quasiharmonic analysis mode to the overall motion can be evaluated by analyzing the eigenvalues of the first 50 modes. The percent contribution of each mode can be determined by taking the reciprocal of the eigenvalue of one mode and dividing by the sum of the inverse eigenvalues for all 50 modes. The eigenvalue is equivalent to the square of the frequency (cm^{-1}). The percent contribution of each mode (Supplemental Figure 4.5) drops off quickly with only the first few modes showing any significant contribution to the overall motion. Modes 1 and 2 in the PXR-RXR heterodimer

and heterotetramer simulations represent proximal percent contributions, while in ERα and PPARγ-RXR simulations mode 2 contributed 50% less to overall motion than mode 1 (Supplemental Figure 4.5). In order to sample the most relevant motions, the first two modes were analyzed for PXR-RXR simulations and only the first mode was analyzed in the ERα and PPARγ-RXR simulations. In all cases, the first mode(s) were sufficient to describe between 18-33% of the overall motion (Supplemental Figure 4.5). To simplify analysis of the PXR-RXR simulations, the sums of the x, y and z vector components of each atom in each mode were obtained and weighted against the percent contribution.

Normal Mode Analysis

Normal mode analysis (NMA) is based on a harmonic approximation of the potential energy function around a minimum energy conformation [41,42]. ELNEMO uses a Hookean potential described by Tirion [41,43], which assumes that the total energy potential function of the reference 3D structure (in this case the crystal structure) is at an energy minimum. In NMA, the lowest energy modes (below 30-100 cm^{-1}) have the largest contribution to the amplitude of atomic displacements. However the first six normal or vibrational modes represent rotational and translational motion and are disregarded [44].

Normal mode theory has been shown to accurately describe large conformational transitions in proteins such as hexokinase [45], lysozyme [46,47] and citrate synthase [48] which occur at microsecond or millisecond time scales. Fifty normal modes were generated using the ELNEMO server for each state of the three nuclear receptors [44]. The only change made was the removal of the modeled loop region (residues 178-197) in the PXR-RXR complexes, as these residues resulted in low frequency modes with low collectivity. Collectivity is a measure of the fraction of residues affected by a given mode. Computed normal modes sometimes have localized motion that corresponds to extended parts of the protein and are usually ignored [44]. This was done to confirm that the high degree of

correlated motion we observed in simulations involving the active-capable forms of nuclear receptors were relevant at longer time scales.

Just as in the quasiharmonic analysis of the all-atom molecular dynamics simulations, we first sought to determine the minimum number of modes required to obtain an accurate description of the overall motion. Supplemental Figure 4.6 shows the percent contribution of each mode, up to the first 50 modes. The first six modes of motion are trivial and have been removed from the analysis. Except for the tetramer, the percent contribution of each of the normal modes appears to drop off more slowly than those of the QH analysis (Supplemental Figures 4.5, 4.6). We chose to analyze modes 7-20, which describe from 48-81% of the overall motion of each nuclear receptor (Supplemental Figure 4.6). To simplify the analysis of the modes, we calculated the vector sum of each atom for modes 7-20, weighted by the percent contribution of each mode.

4.5 ACKNOWLEDGEMENTS

We thank C.D. Fleming, L.M. Guogas, J. Orans, E.A. Ortlund and S. Lujan for helpful advice, P. Wassam, M. Johnson, J. Bischof, and S. Ramachandran for their experimental and computational assistance.

4.6 FIGURE LEGENDS

Figure 4.1. Structural Features of the PXR-RXR Heterotetramer. (A) A model of the PXR (blue, magenta, green)-RXR (yellow) heterotetramer highlights the PXR homodimer interface and the ten-stranded intermolecule β-sheet formed between the two monomers. PXR residues Trp-223 and Tyr-225 central to homodimerization are rendered in yellow with transparent CPK spheres. The □-helices 3, 3', 4 and αAF (green) create the AF-2 surfaces

that bind leucine-rich coactivator peptides like SRC-1 (orange) using a charge clamp (Lys-259, Glu-427) and other residues (light pink). (**B)** Schematics of the oligomeric NR complexes examined in this paper.

Figure 4.2 Conservation of Total Energy During PXR-RXR Simulations. Total energy (kcal/mol), used as a measure of overall simulation stability, remains relatively constant during the course of both the PXR-RXR heterodimer **(A)** and PXR-RXR heterotetramer **(B)** simulations, particularly during the final 10 ns used for analysis (boxed). Both the total energy (grey diamonds) and a running average (black line) are shown.

Figure 4.3. Highly Correlated Motion in the PXR-RXR Heterotetramer. (A) Covariance analysis of the PXR LBD in the PXR-RXR heterodimer and heterotetramer. Residue-residue correlation coefficient values range from blue (anticorrelated, -0.9) to red (correlated, +1), with uncorrelated residue pairs in yellow. Secondary structure is provided from right-to-left, and bottom-to-top. **(B)** Clustering of correlated PXR LBD residues from the PXR-RXR heterodimer simulation. Eleven clusters were identified, five with a correlation cutoff (CC) of 0.6, five with a CC of 0.7, and one with a CC of 0.8. C. Clustering of correlated PXR LBD residues from the PXR-RXR heterotetramer simulation. Three clusters were identified, one each with CCs of 0.6, 0.7, and 0.8. Clusters are colored by the maximum correlation coefficient at which they are observed.

Figure 4.4 Correlated AF-2 Domain Motions in the PXR-RXR Heterotetramer. Vectors describing the motions of PXR LBD □-helices from the heterotetramer **(A)** and heterodimer **(B)** simulations show the active-capable heterotetramer PXR LBD exhibits more overall correlated motion as well as correlation between AF-2 surface helices. Each helix eigenvector (shown by an arrow) is the sum of the α-carbon eigenvectors in that helix. All

arrows were generated using the same scalar magnifications of motion vectors and are presented on the same scale. As such, they represent relative, rather than absolute, movements.

Figure 4.5 Quasiharmonic and Normal Mode Analyses. Angles between motion vectors for all residue pairs in the PXR-RXR heterodimer and heterotetramer. Motion vectors were identified by quasiharmonic analysis (QHA, using the first two modes; **(A)** and by normal mode analysis (NMA, using the first 14 nontrivial modes; **(B)**). In the plots, green represents angles close to zero (correlated), while yellow indicates angles close to 180º (anticorrelated).

Figure 4.6 AF-2 Surface Motions in PPARγ and ERα Complexes. Similar to Figure 4.4, the active-capable PPARγ-RXR heterodimer and ERα homodimer complexes exhibit correlated motions in their AF-2 surfaces during MD trajectories **(A, C)**, while inactive states of both receptors exhibit reduced AF-2 surface correlation **(B, D)**.

4.7 SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 4.1 Conservation of Total Energy During ERα and PPARγ-RXR Simulations. Total energy (kcal/mol), used as a measure of overall simulation stability, remains relatively constant during the course of both ERα and PPARγ-RXR simulations. The final 10 ns (boxed) were used for analysis. Both the total energy (grey diamonds) and a running average (black line) are shown.

Supplemental Figure 4.2 Root Mean Square Deviations from Starting Crystal Structures of PXR LBD Trajectories. Both the all-atom RMSD raw data (grey) and moving

average (black, dashed line; blue, solid line) are plotted for PXR-RXR simulations. Both trajectories have stable RMSDs after approximately 15 ns. The most stable section of the trajectories, 20-30 ns (boxed), was used for analysis.

Supplemental Figure 4.3 Root Mean Square Deviations from Starting Crystal Structures of ERα and PPARγ Simulations. ERα monomer, PPARγ-RXR wild-type, and PPARγ P467L-RXR simulations were stable after 10 ns; data from 10-20 ns (boxed) were used in analysis. The ERα homodimer simulation was stable after 15 ns; data from 15-25 ns (boxed) were used in analysis. Moving averages without raw data are plotted to provide clearer visualization.

Supplemental Figure 4.4 Normalized Covariance Matrices for ERα and PPARγ Simulations. Correlation/anticorrelation versus secondary structure is shown for ERα monomer versus ERα homodimer (**A**) and the PPARγ P467L-RXR mutant heterodimer versus wild-type PPARγ -RXR heterodimer (**B**). Correlation coefficient values are displayed using colors ranging from blue (completely anticorrelated, -0.9) to red (completely correlated, +1) with uncorrelated residue pairs in yellow. Secondary structure is provided from left-to-right and bottom-to-top.

Supplemental Figure 4.5 Percent Contribution to Total Motion by Each Mode of Motion Using Quasiharmonic Analysis (First 50 Modes). Modes 1 and 2 were used to analyze motion in the PXR-RXR simulations; only mode 1 was used for all other simulations.

Supplemental Figure 4.6 Percent Contribution to Total Motion by Each Mode of Motion Using Normal Mode Analysis (First 50 modes). In normal mode analysis, the first six modes of motion are trivial and have been removed from analysis (Methods). Except for

the PXR-RXR heterotetramer, the percent contribution of each of the normal modes appears to drop off more slowly than those in the quasiharmonic analysis. Modes 7-20 were used for analysis, which describe from 48-81% of the overall motion for each nuclear receptor.

Table 4.1

Summary of MD Simulations				
Receptor/Oligomeric State	PDB ID	Length of Simulation		Activity of State
		Total Time	Period Used in Analysis	
PXR-RXR heterodimer*	1ILG	30 ns	20-30 ns	Inactive
PXR-RXR heterotetramer*	1ILG	30 ns	20-30 ns	Active-capable
PPARγ467L-RXR heterodimer	1RDT**	20 ns	10-20 ns	Inactive
PPARγ-RXR heterodimer	1RDT	20 ns	10-20 ns	Active-capable
ERα monomer	1ERE	20 ns	10-20 ns	Inactive
ERα dimer	1ERE	25 ns	15-25 ns	Active-capable
*All PXR simulations are based on 1ILG with residues 178-197 modeled in InsightII.				
**Single-site mutant of PPARγ generated in Pymol. There is no crystal structure of the mutant.				

Table 4.2

θ Angle Analysis of α-carbons of PXR LBD					
$\vec{a}$	$\vec{b}$	QHA		NMA	
		Active-capable PXR from Heterotetramer	Inactive PXR from Heterodimer	Active-capable PXR from Heterotetramer	Inactive PXR from Heterodimer
Lys277 (α4)	Lys259 (α3)	21.0º	70.2º	7.4º	84.5º
Lys259 (α3)	Glu427 (αAF)	47.9º	83.6º	9.7º	104.7º
Lys259 (α3)	Leu424 (αAF)	43.6º	65.1º	23.4º	133.3º
Lys277 (α4)	Glu427 (αAF)	31.5º	66.9º	6.4º	20.6º
Lys277 (α4)	Leu424 (αAF)	32.1º	99.0º	20.2º	48.8º
Leu424 (αAF)	Glu427 (αAF)	12.8º	45.0º	13.9º	28.8º
Average		31.5º	71.6º	13.5º	70.1º

Table 4.3

θ Angle Analysis of α-carbons of PPARγ LBD					
$\vec{a}$	$\vec{b}$	QHA		NMA	
		Active-capable	Inactive	Active-capable	Inactive
		WT PPARγ-RXR	P467L PPARγ-RXR	WT PPARγ-RXR	P467L PPARγ-RXR
Lys277 (α4)	Lys259 (α3)	3.9º	39.4º	29.2º	0.9º
Lys259 (α3)	Glu427 (αAF)	1.8º	40.2º	61.8º	35.8º
Lys259 (α3)	Leu424 (αAF)	1.2º	37.0º	79.6º	104.1º
Lys277 (α4)	Glu427 (αAF)	4.0º	70.1º	44.2º	36.0º
Lys277 (α4)	Leu424 (αAF)	3.0º	49.2º	60.3º	104.1º
Leu424 (αAF)	Glu427 (αAF)	1.3º	26.6º	17.9º	68.7º
Average		2.5º	48.8º	48.8º	58.3º

Table 4.4

θ Angle Analysis of α-carbons of ERα LBD					
		QHA		NMA	
$\vec{a}$	$\vec{b}$	Active-capable ERα Dimer	Inactive ERα Monomer	Active-capable ERα Dimer	Inactive ERα Monomer
Lys277 (α4)	Lys259 (α3)	26.0º	117.8º	47.0º	48.9º
Lys259 (α3)	Glu427 (αAF)	52.1º	94.5º	59.8º	83.4º
Lys259 (α3)	Leu424 (αAF)	47.0º	122.1º	58.1º	96.8º
Lys277 (α4)	Glu427 (αAF)	28.9º	54.1º	30.7º	59.2º
Lys277 (α4)	Leu424 (αAF)	64.0º	39.1º	38.0º	74.4º
Leu424 (αAF)	Glu427 (αAF)	51.4º	28.2º	12.6º	15.6º
Average		44.9º	76.0º	41.0º	63.1º

Figure 4.1

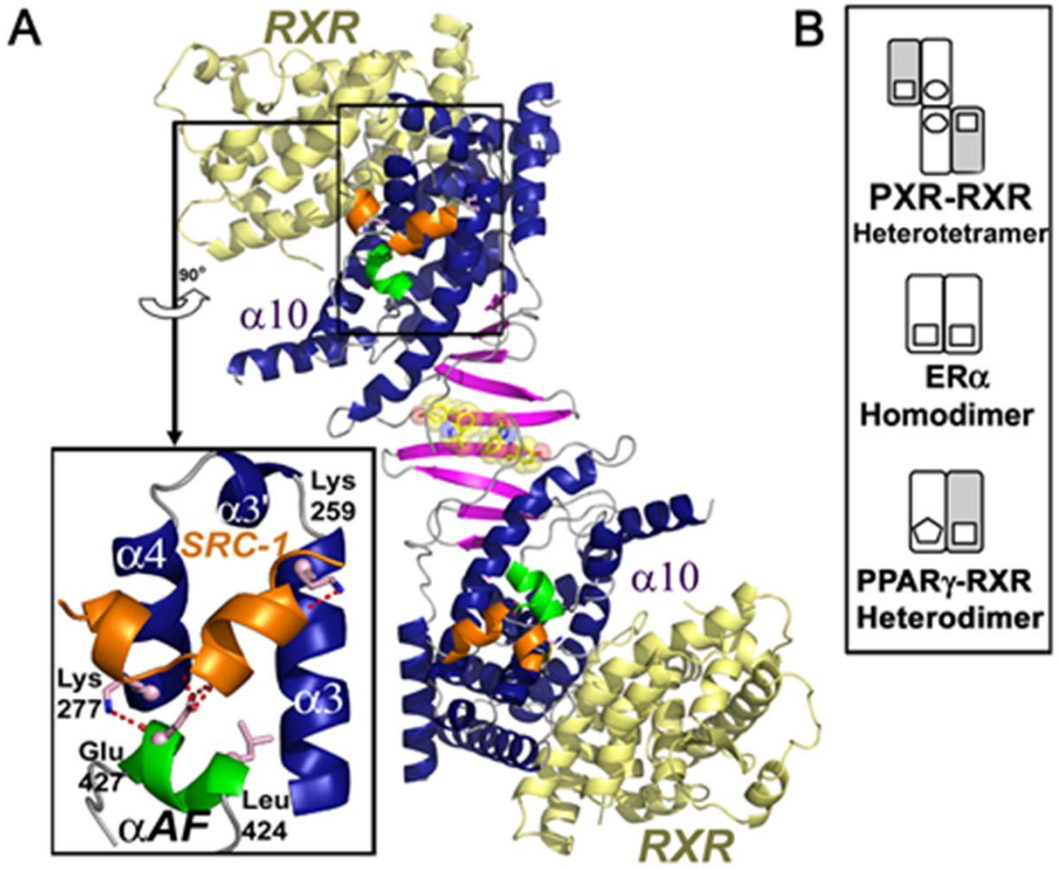

A
RXR
90°
α10
α3
Lys
259
α4
SRC-1
Lys
277
α3
Glu
427
Leu
424
αAF
α10
RXR
B
PXR-RXR
Heterotetramer
ERα
Homodimer
PPARγ-RXR
Heterodimer

Figure 4.2

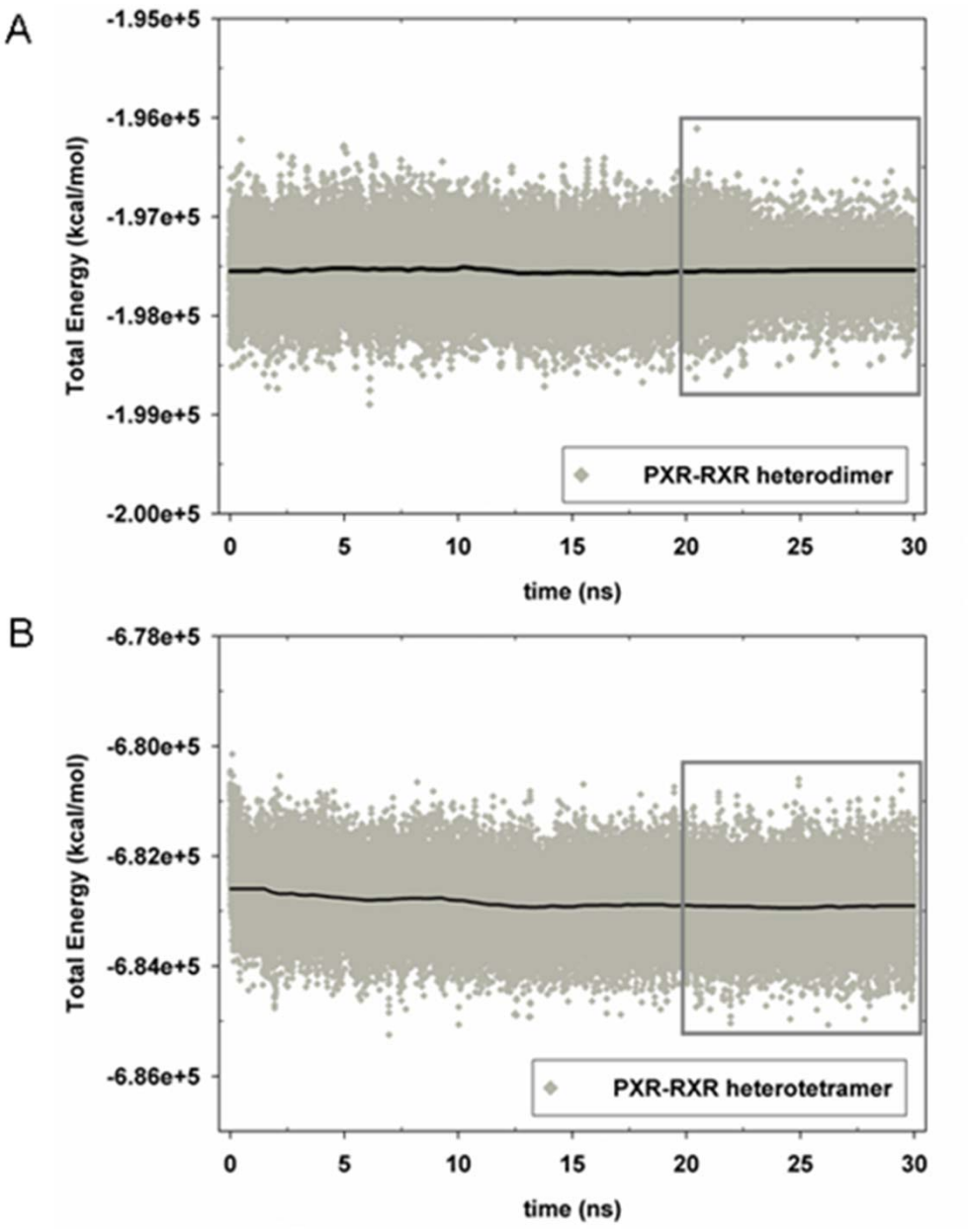

A
-1.95e+5
-1.96e+5
-1.97e+5
-1.98e+5
-1.99e+5
-2.00e+5
Total Energy (kcal/mol)
PXR-RXR heterodimer
0
5
10
15
20
25
30
time (ns)
B
-6.78e+5
-6.80e+5
-6.82e+5
-6.84e+5
-6.86e+5
Total Energy (kcal/mol)
PXR-RXR heterotetramer
0
5
10
15
20
25
30
time (ns)

Figure 4.3

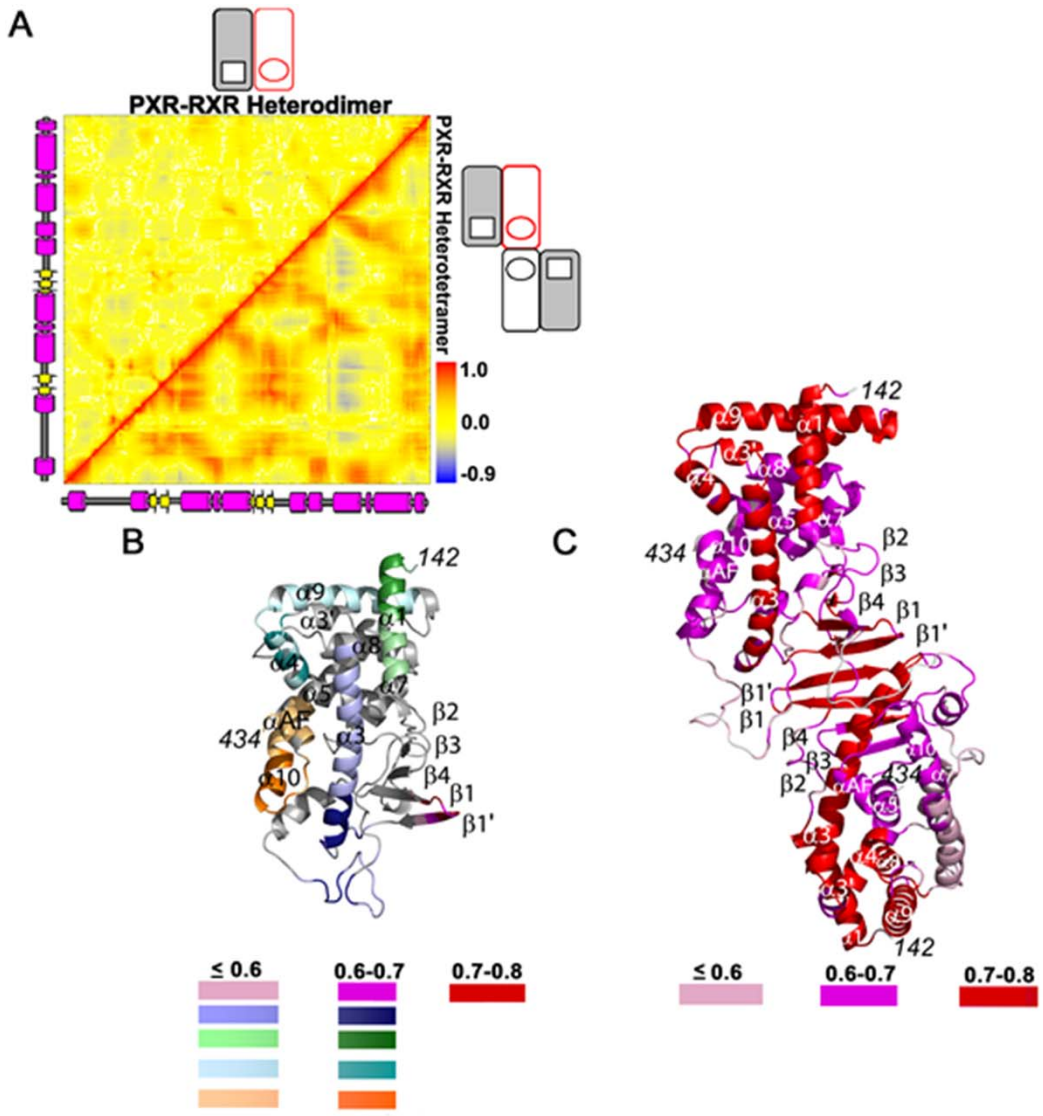

A
PXR-RXR Heterodimer
PXR-RXR Heterotetramer
1.0
0.0
-0.9
B
C
142
434
≤ 0.6
0.6-0.7
0.7-0.8

Figure 4.4

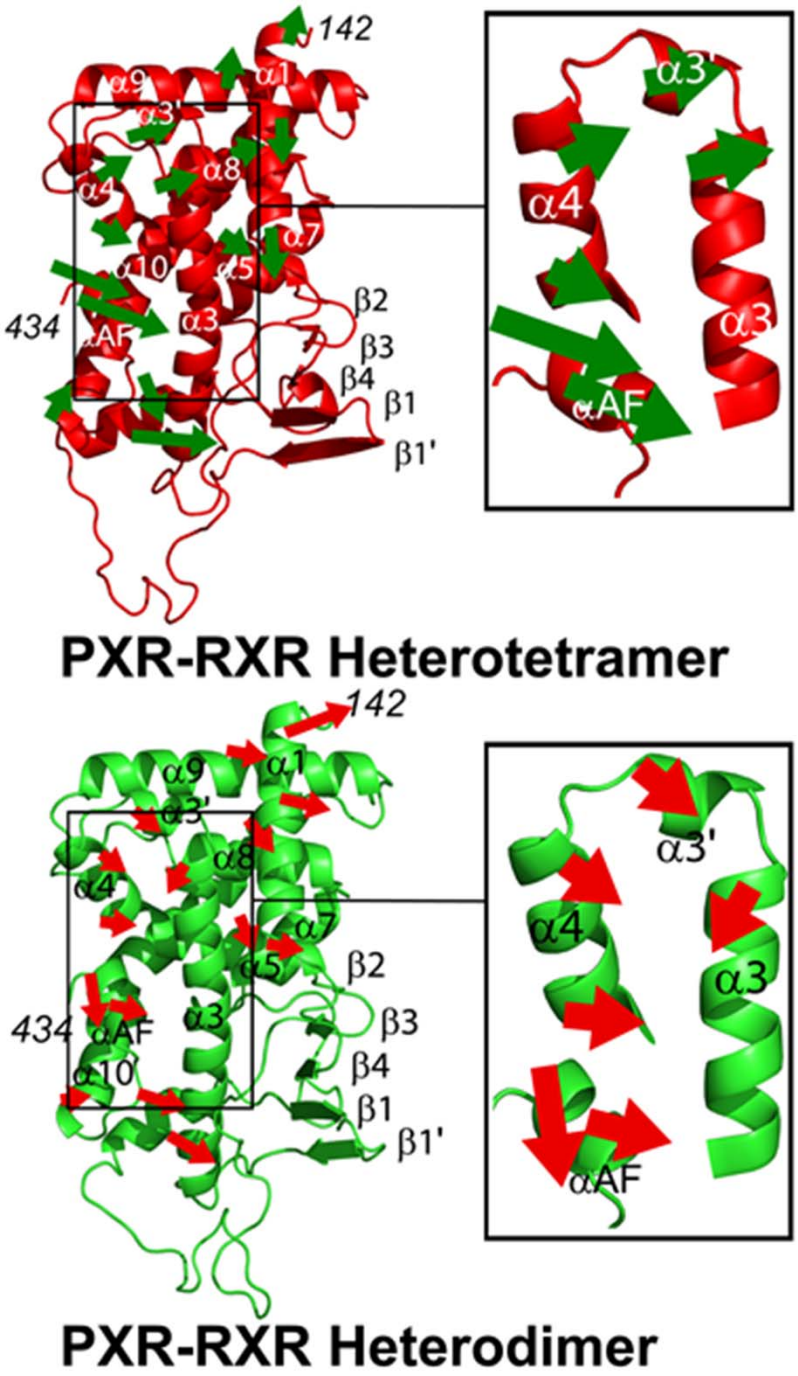

142
α9
α1
α3'
α8
α4
α7
α10
α5
β2
434
αAF
α3
β3
β4
β1
β1'
α3'
α4
α3
αAF
PXR-RXR Heterotetramer
142
α9
α1
α3'
α8
α4
α7
α5
β2
434
αAF
α3
β3
α10
β4
β1
β1'
α3'
α4
α3
αAF
PXR-RXR Heterodimer

Figure 4.5

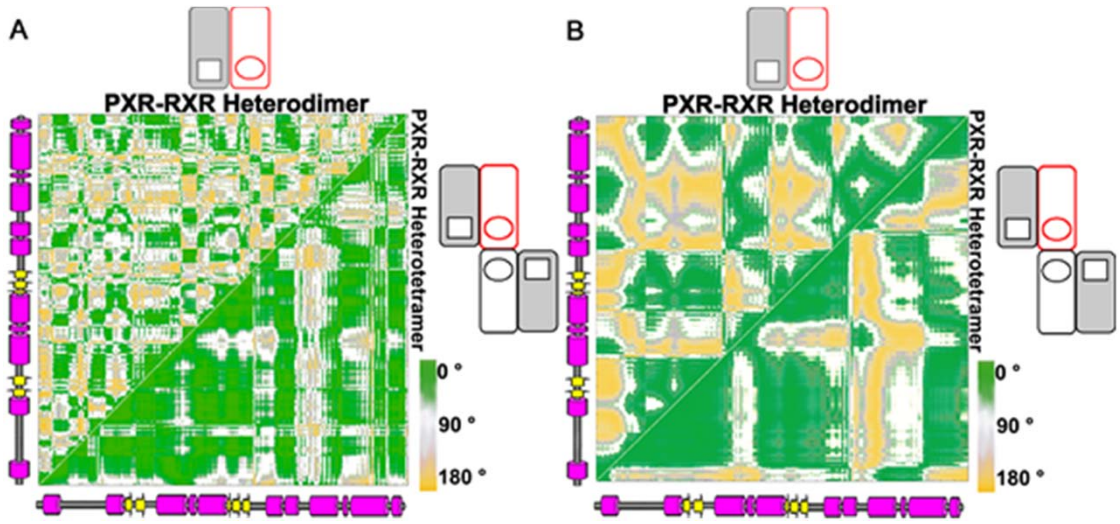
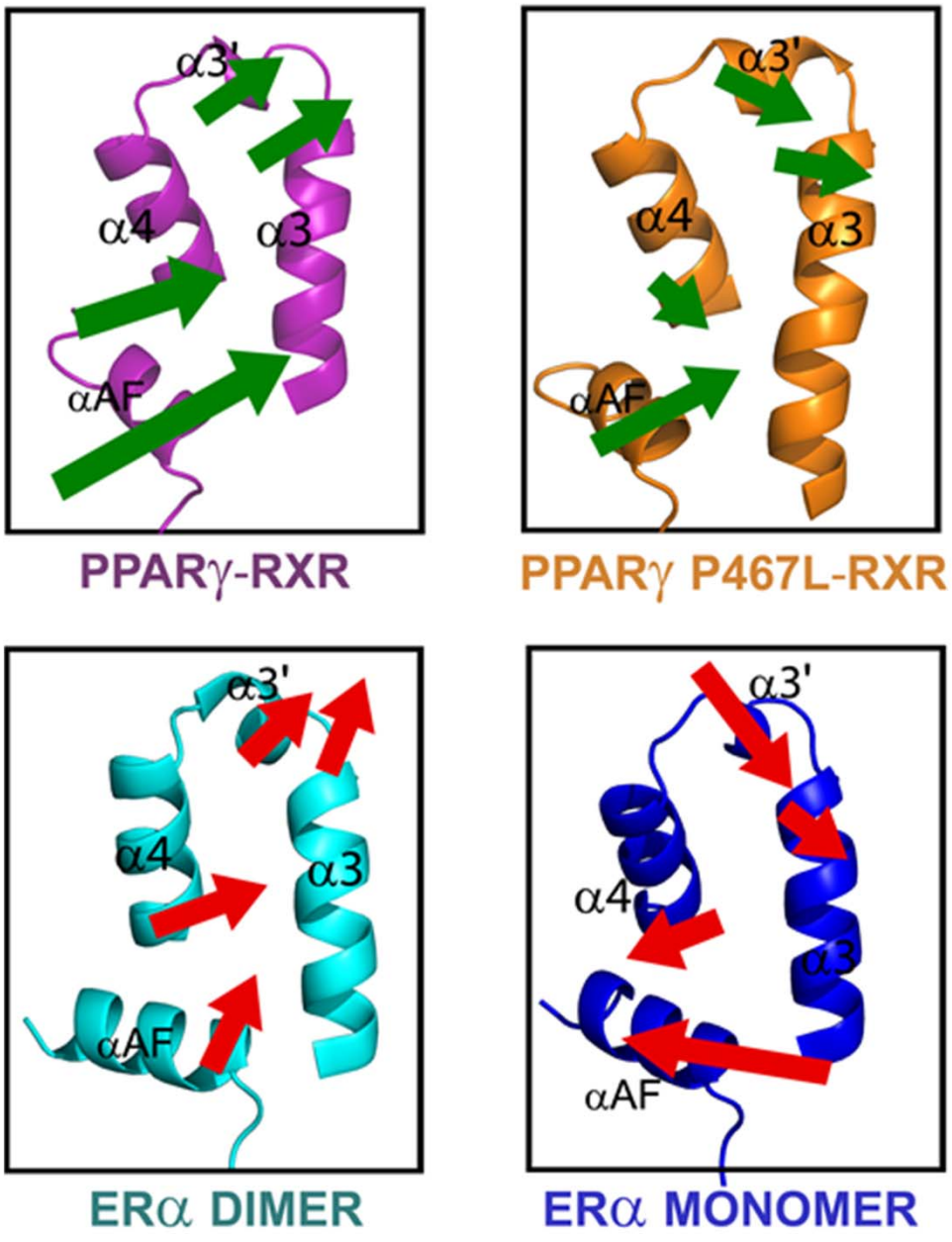
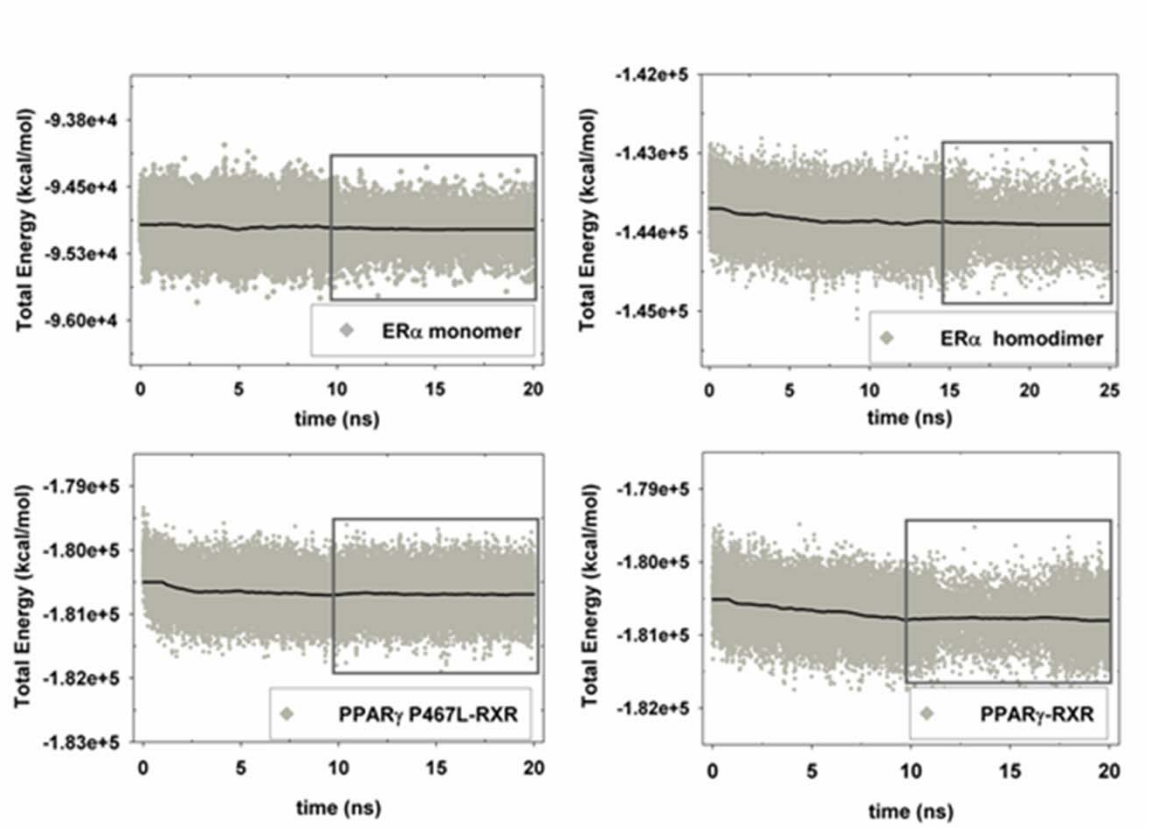

A
PXR-RXR Heterodimer
PXR-RXR Heterotetramer
0 °
90 °
180 °
B
PXR-RXR Heterodimer
PXR-RXR Heterotetramer
0 °
90 °
180 °

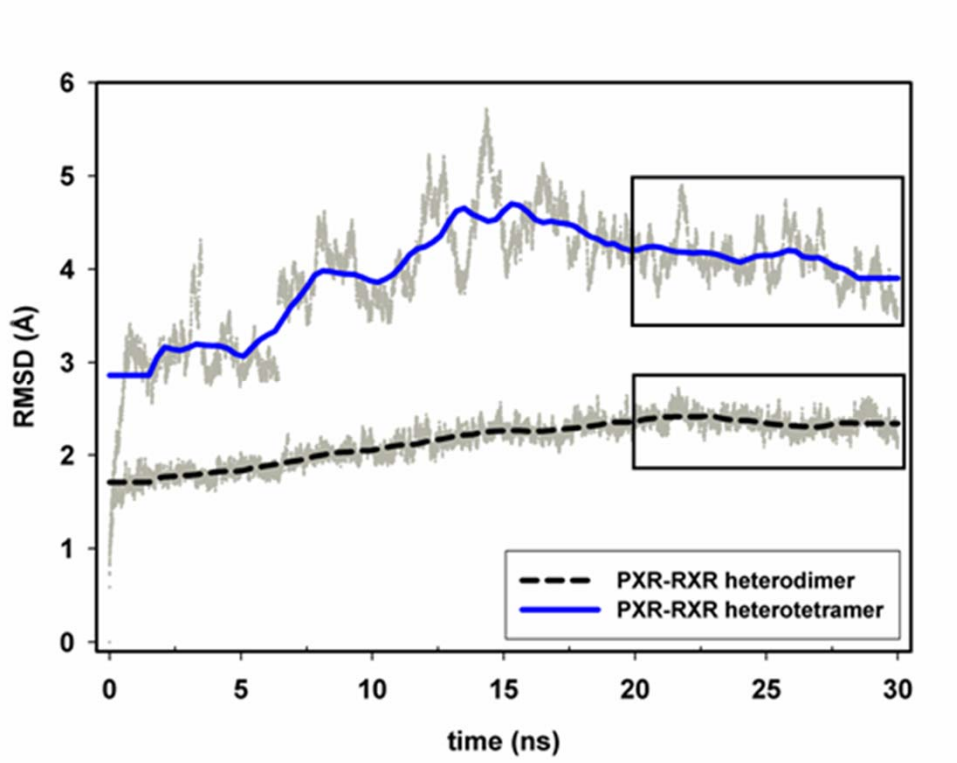
Figure 4.6


α3'
α4
α3
αAF
PPARγ-RXR
α3'
α4
α3
αAF
PPARγ P467L-RXR
α3'
α4
α3
αAF
ERα DIMER
α3'
α4
α3
αAF
ERα MONOMER

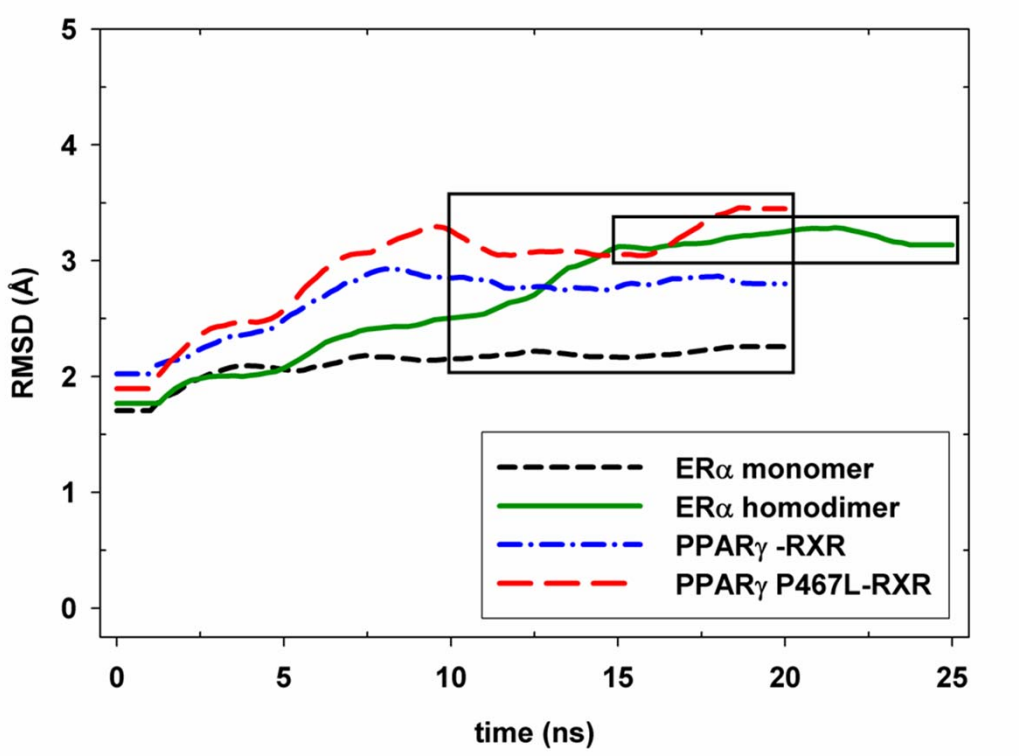
Supplemental Figure 4.1



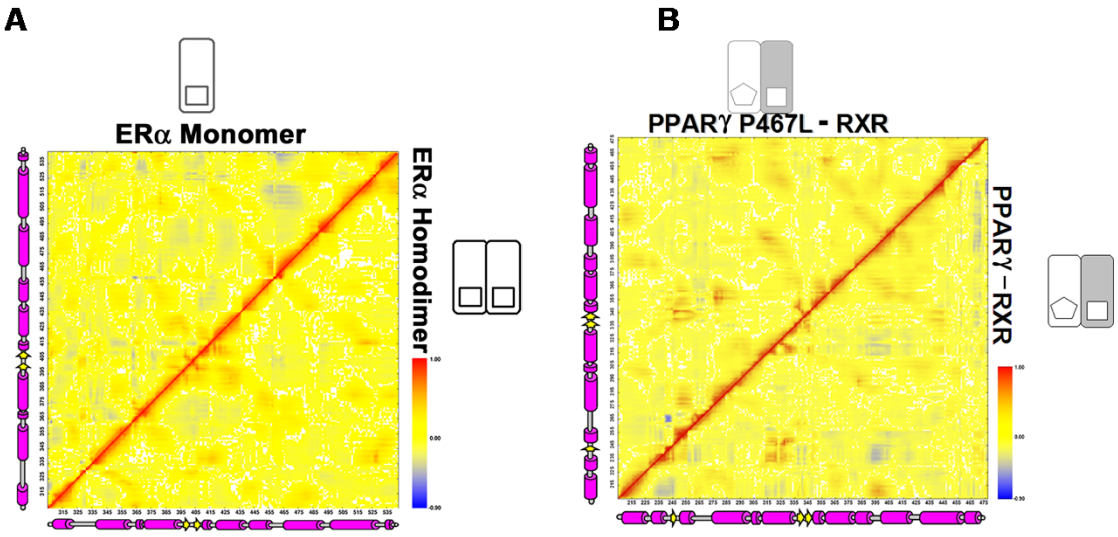
Supplemental Figure 4.2



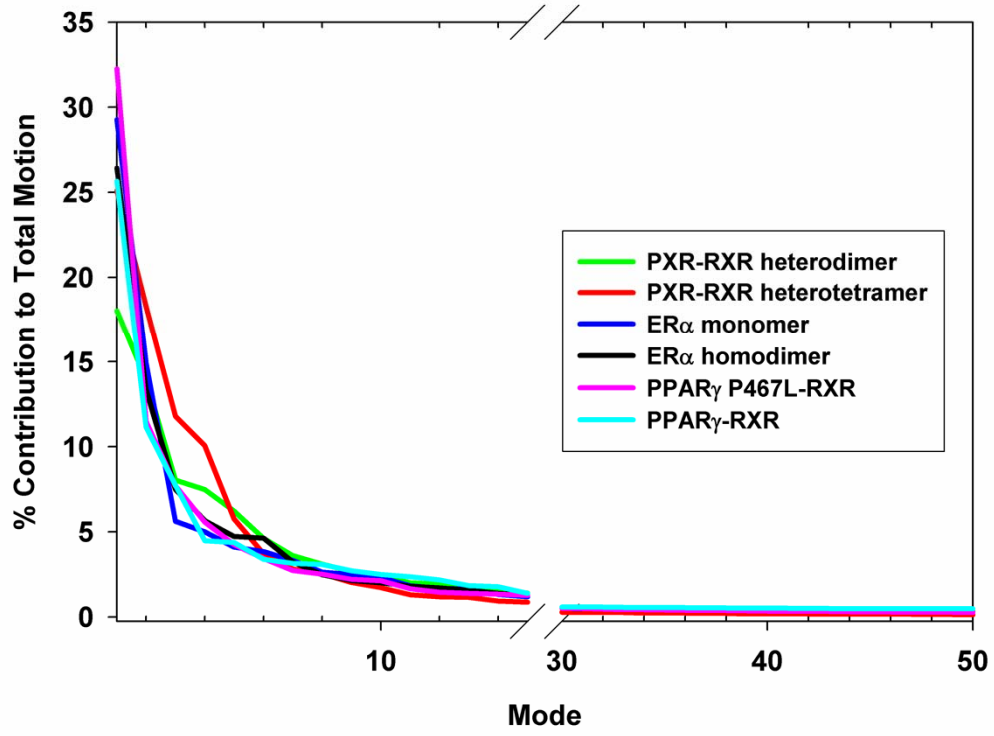
Supplemental Figure 4.3



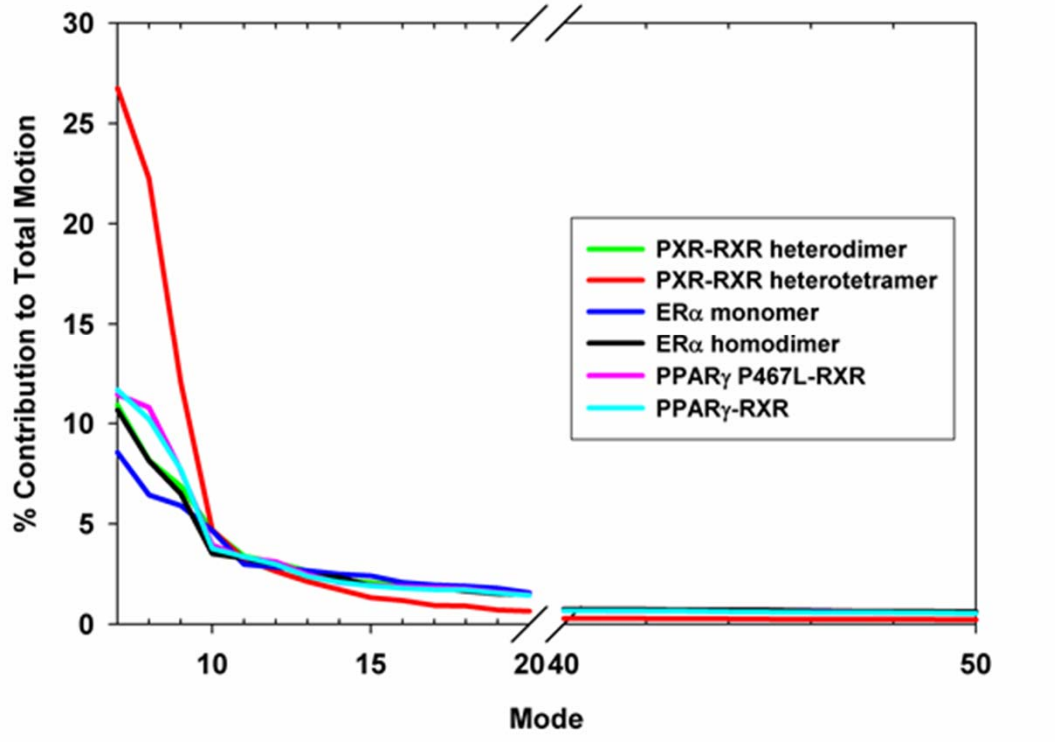
Supplemental Figure 4.4



Supplemental Figure 4.5



Supplemental Figure 4.6



4.9 REFERENCES

1. Staudinger, J., Liu, Y., Madan, A., Habeebu, S., and Klaassen, C. D. (2001) Coordinate regulation of xenobiotic and bile acid homeostasis by pregnane X receptor, *Drug Metab Dispos 29*, 1467-1472.

2. Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engstrom, O., Ohman, L., Greene, G. L., Gustafsson, J. A., and Carlquist, M. (1997) Molecular basis of agonism and antagonism in the oestrogen receptor, *Nature 389*, 753-758.

3. Goodwin, B., Redinbo, M. R., and Kliewer, S. A. (2002) Regulation of cyp3a gene transcription by the pregnane x receptor, *Annu Rev Pharmacol Toxicol 42*, 1-23.

4. Orans, J., Teotico, D. G., and Redinbo, M. R. (2005) The nuclear xenobiotic receptor pregnane X receptor: recent insights and new challenges, *Mol Endocrinol 19*, 2891-2900.

5. Watkins, R. E., Noble, S. M., and Redinbo, M. R. (2002) Structural insights into the promiscuity and function of the human pregnane X receptor, *Curr Opin Drug Discov Devel 5*, 150-158.

6. Krasowski, M. D., Yasuda, K., Hagey, L. R., and Schuetz, E. G. (2005) Evolution of the pregnane X receptor: adaptation to cross-species differences in biliary bile salts, *Mol Endocrinol*.

7. Handschin, C., and Meyer, U. A. (2003) Induction of drug metabolism: the role of nuclear receptors, *Pharmacol Rev 55*, 649-673.

8. Kliewer, S. A., Goodwin, B., and Willson, T. M. (2002) The nuclear pregnane X receptor: a key regulator of xenobiotic metabolism, *Endocr Rev 23*, 687-702.

9. Ekins, S., Mirny, L., and Schuetz, E. G. (2002) A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRalpha, and LXRbeta, *Pharm Res 19*, 1788-1800.

10. Greschik, H., Flaig, R., Renaud, J. P., and Moras, D. (2004) Structural basis for the deactivation of the estrogen-related receptor gamma by diethylstilbestrol or 4-hydroxytamoxifen and determinants of selectivity, *J Biol Chem 279*, 33639-33646.

11. Carlberg, C., and Molnar, F. (2006) Detailed molecular understanding of agonistic and antagonistic vitamin D receptor ligands, *Curr Top Med Chem 6*, 1243-1253.

12. Noble, S. M., Carnahan, V. E., Moore, L. B., Luntz, T., Wang, H., Ittoop, O. R., Stimmel, J. B., Davis-Searles, P. R., Watkins, R. E., Wisely, G. B., LeCluyse, E., Tripathy, A., McDonnell, D. P., and Redinbo, M. R. (2006) Human PXR forms a tryptophan zipper-mediated homodimer, *Biochemistry 45*, 8579-8589.

13. Watkins, R. E., Wisely, G. B., Moore, L. B., Collins, J. L., Lambert, M. H., Williams, S. P., Willson, T. M., Kliewer, S. A., and Redinbo, M. R. (2001) The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity, *Science 292*, 2329-2333.

14. Watkins, R. E., Davis-Searles, P. R., Lambert, M. H., and Redinbo, M. R. (2003) Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor, *J Mol Biol 331*, 815-828.

15. Tamrazi, A., Carlson, K. E., Daniels, J. R., Hurth, K. M., and Katzenellenbogen, J. A. (2002) Estrogen receptor dimerization: ligand binding regulates dimer affinity and dimer dissociation rate, *Mol Endocrinol 16*, 2706-2719.

16. Kliewer, S. A. (2003) The nuclear pregnane X receptor regulates xenobiotic detoxification, *J Nutr 133*, 2444S-2447S.

17. Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993) Essential dynamics of proteins, *Proteins 17*, 412-425.

18. Sharma, S., Ding, F., and Dokholyan, N. V. (2007) Multiscale modeling of nucleosome dynamics, *Biophys J 92*, 1457-1470.

19. Rueda, M., Chacon, P., and Orozco, M. (2007) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics, *Structure 15*, 565-575.

20. Kallenberger, B. C., Love, J. D., Chatterjee, V. K., and Schwabe, J. W. (2003) A dynamic mechanism of nuclear receptor activation and its perturbation in a human disease, *Nat Struct Biol 10*, 136-140.

21. Xue, Y., Chao, E., Zuercher, W. J., Willson, T. M., Collins, J. L., and Redinbo, M. R. (2007) Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism, *Bioorg Med Chem 15*, 2156-2166.

22. Xue, Y., Moore, L. B., Orans, J., Peng, L., Bencharit, S., Kliewer, S. A., and Redinbo, M. R. (2007) Crystal structure of the pregnane X receptor-estradiol complex provides insights into endobiotic recognition, *Mol Endocrinol 21*, 1028-1038.

23. Shulman, A. I., Larson, C., Mangelsdorf, D. J., and Ranganathan, R. (2004) Structural determinants of allosteric ligand activation in RXR heterodimers, *Cell 116*, 417-429.

24. Wang, H., Huang, H., Li, H., Teotico, D. G., Sinz, M., Baker, S. D., Staudinger, J., Kalpana, G., Redinbo, M. R., and Mani, S. (2007) Activated pregnenolone X-receptor is a target for ketoconazole and its analogs, *Clin Cancer Res 13*, 2488-2495.

25. Huang, H., Wang, H., Sinz, M., Zoeckler, M., Staudinger, J., Redinbo, M. R., Teotico, D. G., Locker, J., Kalpana, G. V., and Mani, S. (2007) Inhibition of drug metabolism by blocking the activation of nuclear receptors by ketoconazole, *Oncogene 26*, 258-268.

26. Hustert, E., Zibat, A., Presecan-Siedel, E., Eiselt, R., Mueller, R., Fuss, C., Brehm, I., Brinkmann, U., Eichelbaum, M., Wojnowski, L., and Burk, O. (2001) Natural protein variants of pregnane X receptor with altered transactivation activity toward CYP3A4, *Drug Metab Dispos 29*, 1454-1459.

27. Celik, L., Lund, J. D., and Schiott, B. (2007) Conformational dynamics of the estrogen receptor alpha: molecular dynamics simulations of the influence of binding site structure on protein dynamics, *Biochemistry 46*, 1743-1758.

28. Elhaji, Y. A., Stoica, I., Dennis, S., Purisima, E. O., and Trifiro, M. A. (2006) Impaired helix 12 dynamics due to proline 892 substitutions in the androgen receptor are associated with complete androgen insensitivity, *Hum Mol Genet 15*, 921-931.

29. Renaud, J. P., Rochel, N., Ruff, M., Vivat, V., Chambon, P., Gronemeyer, H., and Moras, D. (1995) Crystal structure of the RAR-gamma ligand-binding domain bound to all-trans retinoic acid, *Nature 378*, 681-689.

30. Johnson, B. A., Wilson, E. M., Li, Y., Moller, D. E., Smith, R. G., and Zhou, G. (2000) Ligand-induced stabilization of PPARgamma monitored by NMR spectroscopy: implications for nuclear receptor activation, *J Mol Biol 298*, 187-194.

31. Chalmers, M. J., Busby, S. A., Pascal, B. D., He, Y., Hendrickson, C. L., Marshall, A. G., and Griffin, P. R. (2006) Probing protein ligand interactions by automated hydrogen/deuterium exchange mass spectrometry, *Anal Chem 78*, 1005-1014.

32. Arnold, L. A., Estebanez-Perpina, E., Togashi, M., Shelat, A., Ocasio, C. A., McReynolds, A. C., Nguyen, P., Baxter, J. D., Fletterick, R. J., Webb, P., and Guy, R. K. (2006) A high-throughput screening method to identify small molecule inhibitors of thyroid hormone receptor coactivator binding, *Sci STKE 2006*, pl3.

33. Arnold, L. A., Estebanez-Perpina, E., Togashi, M., Jouravel, N., Shelat, A., McReynolds, A. C., Mar, E., Nguyen, P., Baxter, J. D., Fletterick, R. J., Webb, P., and Guy, R. K. (2005) Discovery of small molecule inhibitors of the interaction of the thyroid hormone receptor with transcriptional coregulators, *J Biol Chem 280*, 43048-43055.

34. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, *J Comput Chem 24*, 1999-2012.

35. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, and Kollman PA. (2006) AMBER 9, University of California, San Francisco.

36. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, and Klein ML. (1983) Comparison of simple potential functions for simulating liquid water., *J Chem Phys 79*, 926-935.

37. Essman U, Perera L, Berkowitz ML, Darden TA, Lee H, and Pedersen LG. (1995) A smooth particle mesh Ewald method, *J Chem Phys 103*, 8577-8593.

38. Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991) A workbench for multiple alignment construction and analysis, *Proteins 9*, 180-190.

39. Everitt BS, Landau S, and Leese M. (2001) Cluster Analysis. Oxford University Press, Oxford.

40. Brooks, B. R., Janezic, D., and Karplus, M. (1995) Harmonic Analysis of Large Systems. I. Methodology, *Journal of Computational Chemistry 16*, 1522-1542.

41. Tirion, M. M. (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis, *Phys Rev Lett 77*, 1905-1908.

42. Hinsen, K. (1998) Analysis of domain motions by approximate normal mode calculations, *Proteins 33*, 417-429.

43. Tama, F., and Sanejouand, Y. H. (2001) Conformational change of proteins arising from normal mode calculations, *Protein Eng 14*, 1-6.

44. Suhre, K., and Sanejouand, Y. H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement, *Nucleic Acids Res 32*, W610-614.

45. Harrison, R. W. (1984) Variational calculation of the normal modes of a large macromolecule: methods and some initial results, *Biopolymers 23*, 2943-2949.

46. Brooks, B., and Karplus, M. (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme, *Proc Natl Acad Sci U S A 82*, 4995-4999.

47. Gibrat, J. F., and Go, N. (1990) Normal mode analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion, *Proteins 8*, 258-279.

48. Marques, O., and Sanejouand, Y. H. (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations, *Proteins 23*, 557-560.